DOCUMENT RESUME

ED 064 782 EA 004 443

TITLE	An Experiment in Performance Contracting.
INSTITUTION	Office of Economic Opportunity, Washington, D.C. Office of Planning, Research, and Evaluation.
REPORT NO	OEO-Pam-3400-6
PUB DATE	Jun 72
NOTE	236p.
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	*Academic Achievement; *Disadvantaged Youth; *Educational Experiments; Evaluation Methods; Federal Programs; Mathematics Education; *Performance Contracts; Program Evaluation; Reading Skills; School Industry Relationship; *Standardized Tests; Statistical Analysis

ABSTRACT

This report describes the experimental design, presents the contract provisions, and provides conclusions and recommendations. The document is comprised of five chapters that discuss (1) the statistical analysis methods used, (2) the problems of using standardized tests in performance contracting, (3) the contractual procedures between OEO and the 18 school districts and between the school districts and the private firms, (4) an analysis of program costs, and (5) the opinions of school district project managers and those of four of the six participating companies toward the OEO experiment in particular and toward performance contracting in general. A final chapter contains the contractors' statement. The report notes that the results of the experiment indicate that the firms operating under performance contracts did not perform significantly better than did the more traditional school systems. It urges, however, that the results not be interpreted as a blanket finding that educational services and materials should not be purchased under performance-based contracts, nor that private firms cannot provide valuable educational services. This volume includes the more detailed papers that formed the basis for the summary publication (See ED 060 546). (Author/JF)

ED 064782

1-a

Office of
Economic
Opportunity

AN EXPERIMENT IN PERFORMANCE CONTRACTING

EA 004 443

EO Pamphlet 3400 - 6
une 1972

i-b

ED 064782

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

AN EXPERIMENT
IN PERFORMANCE CONTRACTING

Office of Economic Opportunity
Office of Planning, Research, and Evaluation
June, 1972

PREFACE

In February, the Office of Economic Opportunity released a publication, An Experiment in Performance Contracting: Summary of Preliminary Results.[1/] The current volume includes the more detailed papers that formed the basis for that summary publication.

In Chapter I, "A Statistical Analysis of the OEO Experiment in Educational Performance Contracting," Irv Garfinkel and Edward M. Gramlich describe the data used to analyze the experiment, comment on methodological problems confronting analysts, discuss average experiment results for each grade and subject, and summarize major findings on the basis of their own research.

Chapter II, "Implications of Using Standardized Tests in Performance Contracting," by Jeffry Schiller and Ellen Press Murdoch, describes the standardized tests used in this experiment, the process used to select them, scoring techniques, and problems of test reliability and measurement error.

In Chapter III, "Contractual Procedures," Charles Stalford describes the provisions of the contracts between the OEO and the 18 school districts and of the subcontracts between the school districts and the private firms, problems that arose in implementing those provisions during the school year, and adjustments and modifications made to the subcontracts during renegotiation sessions.

[1/] An Experiment in Performance Contracting: Summary of Preliminary Results, OEO Pamphlet 3400-5, February, 1972, Washington, D. C.

In Chapter IV, "Analysis of Program Costs," Mr. Stalford describes the Cost-Ed model developed by Education Turnkey Systems, management support contractor for the experiment, and discusses expenses involved in operating a performance contracting program, both in terms of local costs and of costs adjusted for national averages.

In contrast to the first four chapters, which indicate the views of the OEO staff involved in the experiment, the concluding two sections present the opinions of the local school districts' project managers (Chapter V) and those of four of the six participating educational technology companies (Chapter VI). Both chapters present their authors' views toward the OEO experiment in particular and toward performance contracting in general.

Since the publication of OEO's summary results, much discussion has centered around the specific issues the Agency planned to test and the significance and implications of the experiment's results. The "Summary and Conclusions" section of that paper, therefore, seem worth repeating here:[2/]

[2/] Ibid, pp. 31 and 32.

"In considering the implications of the results presented here, it is important to reiterate what was being tested in the experiment:

"-- The capabilities of a representative group of private education firms using existing instructional materials and technologies and working under specific kind of performance-based contract.

"-- A concept that proponents hoped would be more effective than traditional classroom methods in improving the reading and math skills of poor, under-achieving children.

"The results of the experiment clearly indicate that the firms operating under performance contracts did not perform significantly better than the more traditional school systems. Indeed, both control and experimental students did equally poorly in terms of achievement gains, and this result was remarkably consistent across sites and among children with different degrees of initial capability. On the basis of these findings it is clear that there is no evidence to support a massive move to utilize performance contracting for remedial education in the nation's schools. School districts should be skeptical of extravagant claims for the concept.

"At the same time, the results should not be interpreted as a blanket finding that educational services and materials should not be purchased under performance-based contracts or that private firms cannot provide valuable educational services. Surely performanced based contracts are in some cases a better way to purchase some educational services than the methods currently being used. Surely private firms should continue to play an important role in developing and marketing new educational materials. The results simply say that an uncritical rush to embrace these concepts is unwarranted at this time.

"Some of the benefits of this experiment will not be known for some time, and indeed cannot be precisely pinpointed. The experiment has provoked or added to useful debates on the current use of standardized tests for measuring student performance, on means of introducing change into the educational system, and in general on the subject of accountability. It has raised the possibility that other performers besides schools may sometimes be appropriate providers of education. And hopefully, it will lead to a heightened awareness of the importance of specifying educational goals and measuring progress toward those goals, a process that all too frequently has not been undertaken by school districts.

"But surely the clearest conclusion drawn from the experiment is that we still have no solutions to the specific problem of teaching disadvantaged youngsters basic math and reading skills. Thus while we judge this experiment to be a success in terms of the information it can offer about the capabilities of performance contractors, it is clearly another failure in our search for means of helping poor and disadvantaged youngsters to develop the skills they need to lift themselves out of poverty. The search for solutions to these problems must continue."

Those interested in further details about the experiment or performance contracting in general may wish to consult:

Final Report to the Office of Economic Opportunity: Performance Incentive Remedial Education Experiment (PB 202830), August 31, 1971, Education Turnkey Systems, Inc., Washington, D.C. This report is available for $3.00 from the National Technical Information Service (NTIS), U.S. Department of Commerce, Springfield, Virginia 22151.* This report provides descriptive information on the experiment's operation and an analysis of the costs of the experimental and control programs.

Final Report on the Office of Economic Opportunity Experiment in Educational Performance Contracting (PB 208947), March, 1972, Battelle Memorial Institute, Columbus, Ohio. This report, from the testing and evaluation contractor, also is available for $9.00 from NTIS.* While the OEO analysis deals primarily with overall results, the Battelle report provides detailed site-by-site analyses.

Interim Report on the Office of Economic Opportunity Experiment in Educational Performance Contracting: The Incentives Only Sites, February 7, 1972., Battelle Memorial Institute. This report also will be available from NTIS about June 1, 1972.

A Demonstration of Incentives in Education, OEO Pamphlet 3400-7, February, 1972, Office of Economic Opportunity, Washington, D. C. This and the preceding Battell report discuss the two sites in which private firms were not involved; rather, the instruction was provided under contract with two local teacher associations.

In addition, the Rand Corporation, 1700 Main St., Santa Monica, California 90406, has completed a six volume report, R-900/1-6-HEW, Case Studies in Educational Performance Contracting,* which is available for:

1. R-900/1-HEW, Conclusions and Implications	$3.00
2. R-900/2-HEW, Norfolk, Virginia	$5.00

---

* Copies of these reports are not available from OEO.

3. R-900/3-HEW, Texarkana, Arkansas and Liberty-Eylau, Texas $4.00

4. R-900/4-HEW, Gary, Indiana $4.00

5. R-900/5-HEW, Gilroy, California $3.00

6. R-900/6-HEW, Grand Rapids, Michigan $4.00

Among Rand's five case studies is Grand Rapids, one of the OEO's experiment sites.

TABLE OF CONTENTS

		Page
Preface	....	i
Chapter I.	A Statistical Analysis of the OEO Experiment in Educational Performance Contracting	1
Chapter II.	Implications of Using Standardized Tests in Performance Contracting	51
Chapter III.	Contractual Procedures	109
Chapter IV.	Analysis of Program Costs	159
Chapter V.	Project Managers' Statement	199
Chapter VI.	Contractors' Statement	227

Chapter I

A STATISTICAL ANALYSIS OF THE OEO EXPERIMENT IN EDUCATIONAL PERFORMANCE CONTRACTING

by

Irv Garfinkel

and

Edward M. Gramlich

In the process of preparing this report we have become indebted to a large number of people. We would like to thank our supervisors, John O. Wilson and Thomas K. Glennan, Jr., for getting us started on the project, criticizing our work, and bearing up well when our progress flagged; a review group at the Institute for Research on Poverty at the University of Wisconsin consisting of Arthur Goldberger, Glen Cain, Robert Haveman, and Burt Barnow, for setting us straight on a major error; Fritz Scheuren, Gary Liberson, Jane Lee, and Lester Klein of OEO for statistical advice and computer programming assistance; Jeffry Schiller, Charles Stalford, and Judy Glotzer of OEO for helping us to understand the structure of the experiment and interpret the results; Allen Schenck and Roger Cote of the Battelle Institute for periodic assistance throughout the project; and finally, many other individuals at OEO, too numerous to list, for typing, doing calculations, and criticizing earlier drafts of this paper.

INTRODUCTION

Compensatory education programs have generally failed to improve the cognitive skills of students in need of remedial education. [1] Thus great enthusiasm greeted early reports that a private firm operating under a "performance contract" had succeeded in doubling and even tripling the achievement gains of disadvantaged students in Texarkana, Arkansas. Although the Texarkana project, funded under Title VIII of the Elementary and Secondary Education Act, was intended primarily as a dropout prevention program, the contractual arrangement between the school district and the firm provided that the firm would be paid only to the extent that it improved students' scores on standardized reading and math tests by a prespecified amount. If it failed to meet this standard, it was not reimbursed even for its costs.

Educators, policymakers, and economists alike were intrigued by this attempt to introduce principles of market accountability into the education business. Performance contracting offered the short run promise that the educational technology already accumulated by private firms could be used to improve the cognitive skills of disadvantaged children, and the long run promise that it would encourage innovative firms responding to market incentives to develop educational technology. It would offer the local school board a chance to make decisions on outputs instead of inputs, to select from competing sources of supply of educational services, and to write incentive contracts which might encourage differential focus on certain disadvantaged students, certain subjects, and so forth. As an institution, there was much to be said for it, and there was much initial interest in the Texarkana experience on the part of local school boards.

---

[1] For example, a recent survey of evaluations of compensatory education programs funded by the Office of Education indicated that only 10 of the 1200 were successful in bringing about significant achievement gains.

Although it is impossible to test performance contracting as an institution in any very scientific way, it was possible to take the first step by testing the short run hypothesis that private firms with their already existing technology could outperform the normal public school system in educating disadvantaged students. Accordingly, in the Spring of 1970, the Office of Economic Opportunity decided to run a controlled social experiment in performance contracting. Both experimental and control students in several sites, grades, and subjects were to be given achievement pretests in the Fall of 1970 and post-tests in the Spring of 1971. They and their parents were also to be surveyed to determine family income and structure, parents' education, race, sex; student attendance in the previous and current year; and even parents' attitude towards schools in general and towards innovative programs in particular. At the completion of the post-test the achievement score gains of experimental and control students were to be compared in order to test the hypothesis that the performance contracting firms could outperform the control public schools.

This paper reports on our statistical analysis of these test data. [2] We should emphasize that any evaluative statements we make will be confined to this relatively narrow dimension. It is possible that even this short run experiment in performance contracting had other positive or negative effects on students or schools, but we make no attempt to analyze these other indicators here.

The first section of the paper describes the structure of the experiment and the data we have used for the analysis. The second section contains a detailed discussion of two important methodological and statistical problems which arise in the analysis--the imperfect matching of experimental and control students and measurement error in test scores. In the third section we present and discuss our results for the average experimental effect across all 18 sites, in each grade and subject. The fourth section then disaggregates these results to give individual estimates for each of the eighteen sites, again for each grade and subject. This section also gives some reasons why these individual site results must be interpreted much more cautiously than the overall results. The final section contains a brief summary of the major findings.

---

[2] Battelle Institute was the evaluation contractor for this experiment and they also have a report on it (2). In addition, the Rand Corporation has recently evaluated several other performance contracting experiences (3).

I. Experimental Structure and Data Base

Invitations to participate in the experiment were sent to about 200 school districts which had expressed interest, of whom 163 responded, 77 made a formal application, and eighteen were finally selected. These sites were crudely stratified by size of city and geographical region of the country. Within each site, only elementary and junior high schools which met the criteria for assistance under Title I of the Elementary and Secondary Education Act were chosen to participate in the experiment.

Similar invitations sent to educational technology firms elicited 31 responses. Six firms were finally selected on the basis of corporate experience and interest, the types of achievement they thought they could guarantee, the variety of instructional approach they represented (some firms emphasized hardware and incentives, other curriculum and teacher training methods), and staff qualifications. Each firm was assigned three relatively dissimilar sites.

The companies were to teach disadvantaged students in grades one, two, three, seven, eight, and nine both reading and math for two hours a day in the experimental schools. The performance of these experimental students in reading and math was to be compared with that of similar students in the control schools. To prevent the contractors from "teaching to the tests," which as it turned out was what apparently had happened in Texarkana, experimental students were given separate tests for evaluation and payments purposes, with the evaluation tests (the ones we use here) administered first to prevent practice effects.

Schools and students were assigned to experimental and control groups prior to the pretesting. During the preceding summer, existing reading and math achievement test data from the participating schools in a district were arrayed, with the worst of these schools generally chosen as the experimental school and the second worst the control school. 3/ This ranking procedure was then repeated for students within the two schools. In both the experimental and control schools the 100 lowest ranking students in each grade on this basis were chosen for the experiment. 4/

Some of the students initially included in the experiment moved away during the summer before the experiment and during the experiment itself. In the experimental schools this attrition was replenished from a pool of replacement students in the same school; in the control schools it was not. Usually students who joined or left the experiment in midstream were pre or post-tested at that time, but in order to standardize the analysis, we made no attempt to analyze the test data of these part-time students. We included in our sample only those experimental and control students who were pretested in September and post-tested in June. Rows 1-4 of Table I show that these full year, full test data

---

3/ This condition was violated in some cases because of the presence of other compensatory programs, which would have confounded the results in the worst schools, or because these schools were not willing to participate in the experiment. In addition, two of the districts were so small that the control schools had to be selected from an adjoining district.

4/ The sample size was reduced to 75 for the smaller rural districts to allow them to participate in the experiment.

students represented about seventy percent of the experimental sample and sixty-five percent of the control sample for all grades. The difference in sample size was attributable to the preceding summer's replacement of experimental students. Apart from this quirk, attrition did not seem to affect experimental and control students differentially, and we have no reason to believe that it seriously affects our results.

The selection procedure both for schools and for students suggests that, on average, the control students should have somewhat higher pretest scores than the experimental students. This expectation is confirmed in rows 5-12 of Table I, where we see that both in terms of pretest raw scores (rows 5-8) and grade equivalent conversions (rows 9-12), the control students rank ahead of their experimental counterparts. This fact can also be seen in row 13 of the table, where the correlation coefficient between our experimental dummy variable (which is one for all experimental students and zero otherwise) averages -.14 for the six grades, indicating again that experimental students have somewhat lower pretest scores. Finally, we see from rows 14 and 15 that our sample is also imperfectly matched with respect to average per capita income, which is lower for the experimental students; and race, where the dummy variable indicates that experimental students are more likely to be black. 5/ This imperfect matching of control and experimental students is one of two major problems with these data.

5/ It should be mentioned that we do not have income and demographic data for all students. The response rate for the sex and race of the student is about eighty percent of the full test data sample and that for family income is about fifty five percent. These correlations were each computed for all students where we have the two relevant variables.

TABLE I

Data for the Performance Contracting Analysis By Grade

	Grade 1	Grade 2	Grade 3	Grade 7	Grade 8	Grade 9
Number of Students						
1 . Exp.	1062	1271	1307	1277	1172	1175
2 . Cont.	1083	1135	1156	1153	1128	1005
Percent of Initial Sample						
3 . Exp.	62	74	76	74	68	68
4 . Cont.	63	66	67	67	65	58
Mean Pretest Scores, Reading						
5 . Exp.	70	33	35	40	32	38
6 . Cont.	75	37	42	46	38	45
Mean Pretest Scores, Math						
7 . Exp.	70	28	45	43	39	46
8 . Cont.	75	31	51	48	45	53
Mean Grade Equiv., Reading						
9 . Exp.	NA[a]	1.5	2.2	4.5	4.8	5.6
10. Cont.	NA[a]	1.6	2.3	5.0	5.6	6.4
Mean Grade Equiv., Math						
11. Exp.	NA[a]	1.4	2.2	4.7	5.4	6.0
12. Cont.	NA[a]	1.4	2.3	4.9	5.9	6.6
Correlation of Exp. Dummy Variable with						
13. Pretest Scores	-.10	-.10	-.15	-.12	-.17	-.17
14. Avg. Inc.	-.09	.01	-.06	-.12	-.08	-.11
15. Black Dummy Variable	.12	.09	.13	.09	.14	.04

[a] The first grade pretests were readiness tests for which there are no grade equivalent conversion.

A second major problem, which cannot be inferred from Table I but is nevertheless quite serious, is that a student's achievement test score may not accurately measure his actual achievement level on the day he was tested. The student may not have been feeling well on test day, testing conditions may have been poor in his particular school, he may have cheated or copied answers, or he may have simply made a few lucky guesses. For all of these reasons, we expect some measurement error in test score data.

If measurement errors are random, our results will be biased in a particular, predictable way. In the next section we demonstrate the existence of the bias and present a formula for adjusting the results to eliminate this kind of bias. If, on the other hand, the measurement errors are correlated with experimental status because of poor testing conditions for only one group, the results will be biased in a different way. While there is evidence in test condition reports that in some site-grade-subject combinations measurement errors might be correlated with experimental status--positively in some instances, negatively in others--the reports for most of the sites are too inconsistent and/or incomplete to shed much light on this question. 6/ They do not indicate

---

6/ In some cases the reports contain a qualitative evaluation of the seriousness of reported problems, in some cases there are statements about the percent of students affected by the problems, and in still other cases there is no evaluation of the seriousness of these problems. Moreover, in several instances, apparently serious testing condition problems are reported but there is no indication of what group of students--experimental or control, grade school or junior high--were affected by the problems. See the research report of Battelle Institute (2).

that there is any overall correlation between experimental status and poor testing conditions. Yet they do suggest that for particular site-grade-subject combinations, the assumption that measurement error is random may be untenable. In Section IV, therefore, we present a method for testing the degree to which testing problems unique to control students may be biasing the site-by-site results. But since the method does not allow us to disentangle testing problems unique to experimental students from experimental treatment effects, our site-by-site results must still be interpreted very cautiously.

II. Methodology

In this section we describe the statistical procedures we have used to determine the effect of a policy treatment such as performance contracting in the presence of (a) imperfect matching of experimental and control students; and (b) random measurement error. We first present a brief demonstration of the well-known fact that measurement error in pretest scores biases its regression coefficient towards zero, and the regression constant upwards. We then show that whenever the sample is imperfectly matched, these two statistical problems make it difficult to estimate the true effect of the experiment. Simple comparisons of mean gains of the experimental and control groups will be biased unless the coefficient of true pretest scores is exactly unity. Regression estimates of the experimental effect will be biased by the fact that imperfectly measured pretest scores do not perfectly control for the imperfect matching. Adding other variables to the regression may help reduce the regresion bias, but it would only be an exceptional case where these variables eliminate the bias altogether. Thus there is no simple way to derive an unbiased estimate of the effect of the policy treatment--one must instead try to evaluate the bias directly and then correct the unadjusted estimates.

Let us first assume that achievement levels at post-test time for any student are given by

(1) $\text{POST} = \alpha_0 + \alpha_1 \text{PRE}^* + v,$

where v is a random residual with zero mean and α_o (> 0) and α_1 are the "true" coefficients. Here PRE^* is the unobservable true achievement level at pretest time for this student, or

(2) $PRE = PRE^* + w.$

The residual w is also assumed to have a zero mean and to be completely uncorrelated with post-test scores. Random errors in measuring post-test achievement levels are captured in the v residual of (1).

Since we cannot directly observe true achievement levels, we must estimate this model using observed values for all students

(3) $POST = a_o + a_1 PRE + u.$

This leads to

(4) $a_1 = COV(POST,PRE)/VAR(PRE)$

$a_o = \overline{POST} - a_1 \overline{PRE},$

where $\overline{POST}$ and $\overline{PRE}$ refer to the appropriate means.

But we also know that the true coefficients in (1) are given by

(5) $\alpha_1 = COV(POST,PRE^*)/VAR(PRE^*) = COV(POST,PRE)/VAR(PRE^*)$

$\alpha_o = \overline{POST} - \alpha_1 \overline{PRE}^* = \overline{POST} - \alpha_1 \overline{PRE}.$

Both latter conditions follow from our assumptions about w. Combining (4) and (5) then gives

(6) $a_1 = \alpha_1 p$

$a_o = \overline{POST} - \alpha_1 p \overline{PRE}$

where $p = VAR(PRE^*)/VAR(PRE) < 1$. Measurement error in pretest scores thus biases a_1, the simple coefficient of pretest scores, towards zero, and as a consequence, biases the constant a_o upwards.

We now investigate the effects of these biases on our comparison of experimental and control students. We first note that because students were assigned to experimental and control groups before they were pretested, we can assume that observed pretest scores were generated by the process depicted in (2) for both the experimental and control groups separately. 7/ Thus we let (1) represent the true structural relationship for control students, and we assume that the same structure pertains to the experimental students except that post-test scores are everywhere shifted by α_2, the "true" effect of the experiment. We thus have

$$(7)\ POST^{C} = \alpha_o + \alpha_1\ PRE^{*C} + v$$
$$POST^{E} = \alpha_o + \alpha_2 + \alpha_1\ PRE^{*E} + v.$$

There are several ways in which we could try to measure α_2. The simplest procedure is to compute the simple mean gain differences between the experimental and control students, or

$$(8)\ d = \overline{POST}^{E} - \overline{PRE}^{E} - (\overline{POST}^{C} - \overline{PRE}^{C}).$$

Using the condition that the mean of pretest scores equals the mean of true pretest scores for each group, we can substitute (7) into (8) to derive

$$(9)\ d = \alpha_o + \alpha_2 + (\alpha_1 - 1)\ \overline{PRE}^{E} - \alpha_o - (\alpha_1 - 1)\ \overline{PRE}^{C}$$
$$= \alpha_2 + (\alpha_1 - 1)\ (\overline{PRE}^{E} - \overline{PRE}^{C}).$$

7/ This assumption is not nearly as innocuous as it may seem. If students had been assigned to the experimental and control groups on the basis of observed pretest scores, and the sample imperfectly matched, we could not assume that the within group residuals (w) either had a mean of zero or were uncorrelated with post-test scores. Goldberger (4) discusses this and related points in great detail.

This condition shows that if the mean observed pretest score for control students exceeds the mean experimental pretest score, as it does in our sample, and $\alpha_1 > 1$, the mean gain differences give an estimate of the true experimental effect which is biased against the experimental group. If, on the other hand, $\alpha_1 < 1$, the mean gain differences are biased against the control students. Finally, we note that no matter what is the value of α_1, taking simple mean gain differences is a completely satisfactory estimate of the experimental effect if the sample is perfectly matched.

A second way of estimating the effect of the experiment is to use regression analysis. Typically one would do this if he felt that α_1 was indeed not unity, such that the mean gain differences would give a poor estimate of the experimental effect. But just as a_1 and a_o are both biased in the presence of measurement error, so also is the estimated regression experimental effect. To see this, assume we estimate the model in (3) for the two groups, with the estimated experimental constant now being $a_o + a_2$. The coefficient a_2 would thus be the regression estimate of the effect of the experiment. But from (6) we have

(10) $a_o + a_2 = \overline{POST}^E - \alpha_1 p \overline{PRE}^E$ for experimental students and

$a_o = \overline{POST}^C - \alpha_1 p \overline{PRE}^C$ for control students.

Subtracting and substituting from (7) gives

(11) $a_2 = \alpha_o + \alpha_2 + \alpha_1 (1-p) \overline{PRE}^E - \alpha_o - \alpha_1 (1-p) \overline{PRE}^C$

$a_2 = \alpha_2 + \alpha_1 (1-p)(\overline{PRE}^E - \overline{PRE}^C)$.

Now we see that the regression shift coefficient is a biased estimate of the true effect of the experiment. If the mean pretest score for control students exceeds that of experimental students, which it does here, a_2 will always be biased against the experimental students. The intuitive reason for this is that our regression correction for the imperfectly matched sample, pretest scores, is itself imperfectly measured. Thus our regression coefficient is biased by an expression which is proportional to the product of the two biases. As before, this problem will only arise when the sample is imperfectly matched.

A final way in which we might try to determine the effect of the experiment is to modify the basic regression by including more independent variables. We already know that because of measurement error, observed pretest scores alone will make an insufficient correction for the imperfectly matched sample. Conceivably the other independent variables will improve this correction. Assume for example that instead of (7) we have

(12) $\text{POST} = \alpha_0 + \alpha_1 \text{PRE}^* + \alpha_3 X + v,$

where X is a set of other socioeconomic variables which also influences post-test scores. We show in Appendix I that the expression for the bias in (11) now becomes considerably more complicated, even for only one independent variable, and especially if the mean of X is not the same for experimental and control students. The bias now depends on a set of true and estimated partial correlation coefficients which must be related to one another in a very particular way for the bias to be eliminated entirely. Indeed, we show in the appendix that there is not even a guarantee that including these X variables will reduce, let alone eliminate, the regression measurement error bias.

Thus one must be extremely careful in interpreting experimental results if there is some indication that the sample is imperfectly matched. Taking simple mean gain differences between experimental and control groups will give a biased result except in the unlikely event that the true coefficient of the important independent variable, here pretest scores, is unity. Using an uncorrected regression estimate without other variables will almost certainly lead to a biased result. Using a regression model with other variables included may reduce the bias, but this reduction is by no means guaranteed and it is always possible that including more variables might even make things worse. If α_1 is known or can be estimated, the correct experimental effect can be derived by making appropriate adjustments. But these adjustments, which might sometimes be very complicated, would not be necessary if the sample were perfectly matched.

III. Overall Results

A. Estimation of α_1

Although any direct method we use to estimate the effect of the experiment is likely to be biased, we can correct for these biases if we know α_1, the coefficient of true, unobserved pretest scores. There are two obvious ways in which we could estimate this parameter--either by grouping observations on pretest scores to eliminate measurement error, or by inferring α_1 from separate estimates of a_1 and p. In this section we describe our attempts using both of these methods.

Our previous assumptions suggest that whereas individual scores are made unreliable by measurement error, this measurement error will average out for groups of students. Thus we can eliminate measurement error by aggregating groups of students and then computing α_1 from these aggregations. The trick is to aggregate by groups which are different enough that we can observe points in the pretest-post-test space sufficiently far apart to describe the relationship.

One possible way is the procedure suggested by Wald (8). [8] According to this method, we first rank pretest scores in ascending order for both experimental and control groups for each grade and subject. We then divide both the experimental and control samples in half and compute

$$(13)\ \alpha_1 = \frac{\Delta \text{POST}}{\Delta \text{PRE}} = \frac{\text{POST}^H - \text{POST}^L}{\text{PRE}^H - \text{PRE}^L}$$

---

[8] See also Johnston (4), page 164.

where the H and L superscripts refer to the means of the high and low half-sample respectively. We must do the computations separately for experimental and control groups because if the sample is imperfectly matched, which ours is, pooling all students together will allow treatment effects to confound our estimate of α_1. 9/ Indeed, we could even compute the experimental effect from the difference between these two implied regression lines at the overall pretest mean score, but we prefer to use this estimate of α_1 to adjust mean gain differences or regression coefficients for individual student data. Among other things, adjusting individual site mean gain differences with overall values of α_1 will save us the enormous number of computations necessary in the next section to compute the experimental effect on a site-by-site basis in this way.

A similar averaging technique is suggested by Bartlett (1). This time instead of comparing means of an entire half-sample, we divide both the control and experimental groups into thirds according to pretest score rankings, eliminate the middle third, and compute α_1 from (13) using the means of the highest and lowest groups. This approach will eliminate those students very close to the median, and thereby give students with pretest scores in the tails of the distribution more weight in determining the slope.

9/ Assume for example that our control students are more represented in the upper half-sample and the experimental students in the lower half. If the experiment were very successful,there would be a much tighter distribution of post-test scores than in pretest scores and our estimate of α_1 would be biased downwards.

A third procedure is to aggregate by sites and to estimate (7) directly with these site mean data. In each grade-subject there would be eighteen control observations corresponding to the eighteen control site means, and eighteen experimental site means, which would be included through the use of an experimental dummy variable. The problem with this method is that aggregating by sites may not be a reliable way of eliminating measurement error if one component of this error is testing conditions in an individual site. In this case we would be aggregating a group of students with a residual which does not have an expected value of zero for all students in the group. If this were the case, we would get an estimate of α_1 which is biased by site measurement error.

Table II presents our estimates of α_1 from these three averaging techniques. Apart from the first grade, which is not strictly comparable because the pretest was a readiness test using a different marking scale, all estimates of $\alpha 1$ are nearly equal to or exceed unity. We also note that the Wald and Bartlett procedures (depicted in rows 1-6) give extremely similar results for the experimental, control, and the averaged students. In some cases there are differences between the estimate of α_1 for experimental and control students, but generally even these differences are minor. Finally, the site aggregation technique gives estimates of α_1 which are similar but typically a bit below the sample mean methods. This probably indicates that while site testing problems are present, they may not be too serious.

The second way in which α_1 can be computed is through separate estimates of a_1 and p in equation (6). Estimates of a_1 present no problem for they will be generated in our individual student regressions measuring the effect of performance contracting. We discuss these regressions in more detail in Section IIIB below, and at this point only present the estimates of a_1 in row 8 of Table II.

It is somewhat more difficult to derive estimates of p, which we remember from Section II is equal to $VAR(PRE^*)/VAR(PRE)$. Assume that pretest scores are generated by the process depicted in (2) and that we gave the pretest twice to each student. The relationship explaining separate pretest scores would be

$$(14)\quad PRE^1 = PRE^* + w^1$$
$$PRE^2 = PRE^* + w^2,$$

where the superscripts refer to the first and second test respectively and where w^1 and w^2 are separate independent drawings of the underlying random residual w. As before, w has a mean of zero and is assumed to be completely uncorrelated with true pretest scores.

We could then compute a correlation coefficient for these two observations on pretest scores for each student

$$(15)\quad r = \frac{COV(PRE^1, PRE^2)}{\sqrt{VAR(PRE^1)}\sqrt{VAR(PRE^2)}}.$$

But our assumptions about w indicate that

$$(16)\quad COV(PRE^1, PRE^2) = VAR(PRE^*)$$
$$VAR(PRE^1) = VAR(PRE^2) = VAR(PRE)$$
$$r = \frac{VAR(PRE^*)}{VAR(PRE)} = p.$$

TABLE II

Estimates of α_1

By Grade and Subject

	Grade 1		Grade 2		Grade 3		Grade 7		Grade 8		Grade 9	
	R	M	R	M	R	M	R	M	R	M	R	M
Half Sample[a]												
1. Exp.	.78	.69	1.45	1.00	1.49	1.32	1.13	1.19	1.22	1.13	1.13	1.13
2. Cont.	.81	.75	1.18	.96	1.07	1.14	1.08	1.20	1.13	1.20	1.01	1.12
3. Avg.	.80	.72	1.32	.98	1.28	1.23	1.11	1.19	1.18	1.16	1.07	1.13
Third Sample[b]												
4. Exp.	.78	.68	1.46	1.03	1.48	1.31	1.13	1.19	1.23	1.14	1.13	1.15
5. Cont.	.82	.74	1.16	.97	1.08	1.14	1.09	1.20	1.13	1.22	1.02	1.11
6. Avg.	.80	.71	1.31	1.00	1.28	1.23	1.11	1.19	1.18	1.08	1.08	1.13
7. Site Agg.[c]	.72	.53	1.25	1.12	1.17	1.27	1.07	1.19	1.11	1.15	1.05	1.13
8. Estimate of a_1[d]	.48	.38	.90	.71	.92	.96	.95	.96	.89	.96	.88	.87
9. Estimate of α_1; computed reliability coefficients[e]	.51	.40	.98	.81	.98	.98	1.02	1.03	.98	1.06	.95	.93
10. Estimate of α_1; pre, post correlations[f]	.76	.62	1.36	1.04	1.25	1.26	1.10	1.14	1.07	1.14	1.02	1.09

---

[a] **Computed by dividing the sample in half and using the formula**

$$\alpha_1 = \frac{\overline{POST}^H - \overline{POST}^L}{\overline{PRE}^H - \overline{PRE}^L}$$ **, where the superscripts H and L refer to the high and low halves of the sample.**

[b] **Computed by dividing the sample into thirds, eliminating the middle third, and using the formula**

$$\alpha_1 = \frac{\overline{POST}^H - \overline{POST}^L}{\overline{PRE}^H - \overline{PRE}^L}$$

[c] Computed from the regression

$\overline{POST}_j = \alpha_o + \alpha_1 \overline{PRE}_j + \alpha_2 EXP$, where EXP is an experimental dummy variable (1 for experimental students) and where $\overline{POST}_j$ and $\overline{PRE}_j$ refer to the mean post-test and pretest scores for the j^{th} site.

[d] Evaluated at the mean pretest score in the regression $POST = a_o + a_1 PRE + a_2 EXP + a_3 PRE^2 + a_4 PRE^3 + \sum_{i=1}^{17} a_4 + {}_i SITE$.

[e] Computed by dividing a_1 by p_1 where a_1 is as given in row 8 and p is the computed Kuder-Richardson reliability coefficient for the pretest achievement tests.

[f] Computed by dividing a_1 by dividing a_1 by r, where a_1 is as given in row 8, and r is the pre-post correlation coefficient.

Thus this correlation coefficient would give us a direct estimate of p.

Our first such estimate comes from the reliability coefficients computed by Battelle Institute for these tests. These reliability coefficients, calculated from the Kuder-Richardson formula #21, measure the internal consistence of test answers. 10/ Since they do not account for random measurement errors due to variations in testing conditions and day to day variations in individual student performance, they will generally over-estimate p and underestimate α_1 (a_1/p). We see from row 10 of Table II that the estimates of α_1 derived in this way are indeed below our previous estimates--by about .25 for the lower grades and .15 for the upper grades. They are close to the estimate of a_1 in row 8 because the reliability coefficients are close to unity.

We can derive an alternative estimate of p from the simple pretest, post-test correlations in our own sample. These correlations will be less than unity not only because of measurement errors which are attributable to imperfect test instruments, imperfect test conditions, and abnormal student performances, but they will also be less than unity because of true changes in achievement level, including those resulting from performance contracting, which took place during the school year. Since true changes in achievement level as well as measurement error will reduce the PRE-POST correlation, these correlations give us a lower bound to p and an upper bound to α_1.

10/ See Saupe (7) and the Battelle final report (2).

Our estimates of α_1 using this method are given in row 10 of Table 11. As expected, they are higher than those using the computed reliability coefficients. But they are on the whole quite close to the estimates of α_1 derived by the averaging methods, sometimes even slightly below these previous estimates. This seems to indicate that, at least for our sample, the estimates of p computed in this way are more realistic than those implied by Kuder-Richardson statistics. Since all sets of estimates but those in row 9 are quite consistent, it scarcely matters which of these sets we use for adjustment purposes, and we have arbitrarily chosen those in row 3.

B. The Effect of the Experiment

We now turn to our estimates of the overall effect of the experiment. We have computed these estimates in the manner discussed above--first by mean gain differences, then by a regression analysis without other independent variables, and finally by a regression analysis with other independent variables included. All of these estimates, both unadjusted and adjusted for the biases of Section III, are given in Table III.

The first row of Table III gives the mean gain differences computed exactly as in (8), or $d = (POST^E - PRE^E) - (POST^C - PRE^C)$. The second row gives the regression estimates of a_2 from the model

$$(17)\ POST = a_o + a_1 PRE + a_2 EXP + a_3 PRE^2 + a_4 PRE^3 + \sum_{i=1}^{17} a_{4+i} SITE_i,$$

where EXP is an experimental dummy variable. In addition to the basic specification discussed above, we have added a square and cubic term to adjust for possible nonlinearities in pretest scores, and seventeen site dummies (the intercept for the eighteenth site is the overall constant, a_o). The third row gives regression estimates of a_2 from the model

$$(18)\ \text{POST} = a_o + a_1\text{PRE} + a_2\text{EXP} + a_3\text{PRE}^2 + a_4\text{PRE}^3 + a_5\text{X}$$

$$+ \sum_{i=1}^{17} a_5 + {}_i\text{SITE}_i,$$

where X is a vector of other independent variables including average family income, education of parents, race, sex and age. 11/ (The full regressions used here and in the next section are available on request.) The fourth row presents the mean gain differences from the first row adjusted by expression (9), which is the bias arising whenever

$\alpha_1 \neq 1$, or

$$(19)\ \alpha_2 = d - (\alpha_1 - 1)(\overline{\text{PRE}}^E - \overline{\text{PRE}}^C).$$

The fifth row presents the adjusted regression coefficients derived by adjusting the coefficients in the second row by expression (11), or

$$(20)\ \alpha_2 = a_2 - \alpha 1\ (1\text{-}p)(\overline{\text{PRE}}^E - \overline{\text{PRE}}^C).$$

The most striking aspect of Table III is the small size of the differences between experimental and control students. The largest difference anywhere in the Table is 3 raw score points, which converts to about .3 grade equivalent units in the eighth grade and less in the other grades. The small differences are seen in both reading and in math, in lower and upper grades, and by all five unadjusted and adjusted methods.

Examining the Table in more detail, we see first that the mean gain differences and unadjusted regression estimates are rather close together, though the regression gives a more negative picture in every

---

11/ Observations with missing demographic data were assigned the mean values in the sample.

TABLE III

Difference in Achievement Test Raw Scores Between Experimental and Control Students By Grade and Subject

	Grade 1		Grade 2		Grade 3		Grade 7		Grade 8		Grade 9	
	R	M	R	M	R	M	R	M	R	M	R	M
1. Mean Gain Differences[a]	1	0	-1	1	2	0	1	-1	-1	-2	1	0
2. Regression Without Other Variables[b]	-2	-3*	-2*	0	0	-1*	0	-2*	-1*	-3*	0	0
3. Regression With Other Variables[c]	-1	-3*	-2*	-1	-1	-2*	0	-1*	-1*	-3*	0	0
4. Adjusted Mean Gain Differences[d]	0	-1	0	1	3	1	2	0	0	-1	1	1
5. Adjusted Regression Coefficients[e]	0	-1	0	1	2	1	1	-1	1	-2	1	1

---

[a] Showing $d = \overline{POST}^E - \overline{PRE}^E - (\overline{POST}^c - \overline{PRE}^c)$.

[b] Showing a_2 in $POST = a_o + a_1PRE + a_2EXP + a_3PRE^2 + a_4PRE^3 + \sum_{i=1}^{17} a_{4+i}SITE_i$.

[c] Showing a_2 in $POST = a_o + a_1PRE + a_2EXP + a_3PRE^2 + a_4PRE^3 + a_5X + \sum_{i=1}^{17} a_{5+i}SITE_i$.

[d] Showing $\alpha_2 = d - (\alpha_1 - 1)(\overline{PRE}^E - \overline{PRE}^C)$, where d is as in row 1.

[e] Showing $\alpha_2 = a_2 - \alpha_1(1-p)(\overline{PRE}^E - \overline{PRE}^C)$, where a_2 is as in row 2.

*Indicates statistical significance at .05 level.

case. That the regression estimate would tend to be more negative follows from our earlier discussion which indicated that the unadjusted regression coefficients would be biased against the experimental students by α_1 (1-p) of the difference in pretest means, where α_1 (1-p) averages about .26, while the mean gain differences would be biased by (α_1-1), or generally about .12, of the difference in pretest means. As we have shown, adjusting both appropriately as in rows 4 and 5 brings the two methods quite close together. What may be more surprising is that the regression estimates with the other variables included are so little different than the unadjusted regression coefficients. These other variables do not appear to reduce the measurement bias.

It is of course possible that these unimpressive overall relative gains could be attributed to especially good performance on the part of the control students. If this were so, we would observe high absolute gains for experimental and control students alike. Rows 1-3 of Table IV indicate, however, that this is not the case for the experimental students. The grade equivalent gains in row 3 are uniformly less than 1.0, especially in the lower grades. This is much less than the companies had predicted and implies that these students will fall even further behind their pre-experimental levels. [12]

We can also look at these gains in a different light. Rows 4 and 5 of Table IV compare the start and end of the year position of the experimental students. In the second grade, for example, students pretest

---

[12] A Rand Corporation study of several performance contracting programs other than those in the OEO experiment arrives at similar conclusions. See (3).

TABLE IV

Experimental Students Mean Pretest Scores, Mean Post Test Scores and Means Gains in Grade Equivalents By Grade and Subject

	Grade 1		Grade 2		Grade 3		Grade 7		Grade 8		Grade 9	
	R	M	R	M	R	M	R	M	R	M	R	M
1. Pretest Score	NA	NA	1.5	1.4	2.2	2.2	4.5	4.7	4.8	5.4	5.6	6.0
2. Post Test Score	1.0	1.3	1.9	1.9	2.5	2.6	4.9	5.3	5.7	6.2	6.4	6.8
3. Gain	NA	NA	.4	.5	.3	.4	.4	.6	.9	.8	.8	.8
4. Grade Behind at Start[a]	NA	NA	.5	.6	.8	.8	2.5	2.3	3.2	2.6	3.4	3.0
5. Grade Behind at End[b]	.9	.6	1.0	1.0	1.4	1.3	3.0	2.6	3.2	2.7	3.5	3.1

---

[a] For any grade subtract the pretest mean from the grade number.

[b] For any grade subtract the post-test mean from the next highest grade number less .1

at a level of 1.5 in reading and 1.4 in math, which indicates that they are .5 and .6 grades respectively behind their proper level (row 4). If the experiment were successful, ending first grade students would be well ahead of these levels. But in fact they are not. As is indicated in row 5, experimental students ending first grade have deficiencies of .9 and .6 respectively, and these patterns are repeated for the other three grades where such comparisons are possible.

Finally, we investigate the possibility that performance contracting differentially affected students at different points in the pretest distribution. It could be, for example, that the incentive structure used for the performance contracting companies encourages them to concentrate more on the better students or more on the worse students. If the former were the case, the coefficient of pretest scores would be higher for experimental students; otherwise it would be lower. To examine this possibility, we have added a slope dummy variable to our basic regression model

$$(21)\ \mathrm{POST} = a_o + a_1\mathrm{PRE} + a_2\mathrm{EXP} + a_3\mathrm{PRE}^2 + a_4\mathrm{PRE}^3 + a_5(\mathrm{EXP})(\mathrm{PRE}) + \sum_{i=1}^{17} a_{5+1}\mathrm{SITE}_i .$$

A negative value of a_5 indicates that the worst students fared relatively better in the performance contracting schools.

The results of this test for a_2 and a_3 are given in Table V. Since we are only interested in the board question of whether there is differential improvement, we have not adjusted these simple regression coeffi-

TABLE V

Effects of Performance Contracting on Students
at Different Pretest Levels
By Grade and Subject[a]

	Grade 1		Grade 2		Grade 3		Grade 7		Grade 8		Grade 9	
	R	M	R	M	R	M	R	M	R	M	R	M
Estimate of 1. a_2	3	4*	-2	0	0	-2	1	1	-1	0	0	1
2. a_5	-.06*	-.09*	.01	0	0	.01	-.02	-.06*	-.02	-.08*	0	-.04

[a] From $\text{POST} = a_o + a_1\text{PRE} + a_2\text{EXP} + a_3\text{PRE}^2 + a_4\text{PRE}^3 + a_5(\text{EXP})(\text{PRE}) + \sum_{i=1}^{17} a_5 + {}_i\text{SITE}_i.$

*Indicates statistical significance at .05 level.

cients for measurement bias. Thus we cannot say whether they indicate that experimental or control students are better off, or if so by how much. But we can see that the difference in slope between experimental and control students is very slight, being statistically significant in only a few cases and never amounting to much quantitatively.

Thus no matter how we look at these overall results, whether by comparing experimental and control students, adjusting or not adjusting for bias, looking at absolute or relative gains, or testing for differential improvements between better and worse students, we cannot find significant experimental effects. The performance of the performance contractor was, on average, no better than the performance of the control public schools.

IV. Site-by-Site Results

In this section we examine the possibility that our overall neutral results can be explained by offsetting successes at some sites and failures at others. Before proceeding with an examination of the data, however, we should note that these individual site results are much less reliable than our overall conclusions. For one thing, the sample size at any particular site is obviously much smaller than the overall sample size. Second, and possibly related to the first reason, careful inspection of the site-by-site results suggests that at some sites, experimental-control differences might have resulted from extraordinarily large or small gains of the control rather than the experimental group. Third, and also related, results at any particular site are likely to be far more sensitive to testing conditions, which were at times less than ideal, than the overall results.

Tables VI-1 to VI-6 present these overall site-by-site comparisons. The first two columns give the simple mean gain differences as before. The second two columns follow the regression formulation in (17), except that now the overall experimental dummy is dropped and eighteen individual dummies are substituted in its place to give different experimental effects in each site. The fifth and sixth columns add the other independent variables as in (18), again with eighteen experimental dummies. The seventh and eighth columns then give the mean gain differences adjusted by expression (19). As was shown above, these numbers are virtually identical to the adjusted regression coefficients. And finally, the

Table VI-1

Difference in Achievement Test Raw Scores Between Experimental and Control Students by Site, Grade and Subject - Grade 1

	Mean Gain Difference [a]		Regression Model III [b]		Regression Model IV [c]		Adjusted Mean Gain Difference [d]		Adjusted Pooled Mean Gain Difference [e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer	14*	5	12*	4	12*	4	14	4	8	5
Athens	7*	8*	-1	-	-1	1	4	4	3	6
Wichita	13*	-2	-15*	-3	-14*	-2	-13	-2	-5	-5
Dallas	9*	6	-3	0	-3	-1	8	6	17	13
Anchorage	5*	10*	-1	0	2	2	2	6	-4	-6
Rockland	0	3	-4	-3	-4	-2	-2	0	-3	-1
Las Vegas	0	4	-4	-12*	-5*	-12*	-2	-8	-1	-3
Fresno	-1	11*	1	- 9*	2	-9*	0	-11	-12	-17
Philadelphia	-12*	-17*	-7*	-11*	-6*	-10*	-10	-14	1	-7
Taft	3	1	-	-3	0	-2	1	-1	16	11
Grand Rapids	9	-	-4	-7*	-7*	-8*	6	-4	6	-3
Hartford	-4	-10*	-5*	-9*	-5	-9*	-2	-10	-5	-3
McComb	14*	25*	-	10*	3	12*	8	17	11	16
Seattle	10*	-1	-13*	-9*	-12*	-8*	13	-5	-5	-7
Portland	-4	-10*	-9*	-15*	-10*	-15*	-6	-12	-6	-8
Jacksonville	13*	9*	21*	14*	20*	15*	14	11	4	4
Hammond	-7*	2	-11*	-2	-11*	-2	-8	0	-2	0
Bronx	NA	NA	NA	NA	NA	NA			NA	NA

Footnotes for Table VI

a. Showing $d_j = (\overline{POST}^E - \overline{PRE}^E)_j - (\overline{POST}^C - \overline{PRE}^C)_j$, $j = 1,18$.

b. Showing a_{21+j} in $\overline{POST} = a_o + a_1 PRE + a_3 PRE^2 + a_4 PRE^3 + \sum_{i=1}^{17} a_{4+i} SITE_i + \sum_{j=1}^{18} a_{21+j} (EXP)(SITE_j)$.

c. Showing a_{22+j} in $POST = a_o + a_1 PRE + a_3 PRE^2 + a_4 PRE^3 + a_5 X + \sum_{j=1}^{18} a_{22+j} (EXP)(SITE_j)$.

d. Showing $\alpha_{2j} = d_j - (\alpha_1 - 1)(PRE^E - PRE^C)_j$ where d_j is as in column a.

e. Showing $\alpha_{2j} = d_j^* - (\alpha_1 - 1)(\overline{PRE}^E - \overline{PRE}^C)_j$, where $d_j^* = (\overline{POST}^E - \overline{PRE}^E)_j - (\overline{POST}^C - \overline{PRE}^C)$, $j = 1,18$.

*Indicates statistical significance at .05 level.

Table VI-2

Difference in Achievement Test Raw Scores Between Experimental and Control Students by Site, Grade and Subject - Grade 2

	Mean Gain Difference[a]		Regression Model III[b]		Regression Model IV[c]		Adjusted Mean Gain Difference[d]		Adjusted Pooled Mean Gain Difference[e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer, Tennessee	2	2	5*	2	5*	2	-1	2	2	0
Athens, Georgia	1	0	-1	0	-2	-1	2	0	0	0
Wichita, Kansas	9*	1	1	-3*	0	-3*	13	1	7	-1
Dallas, Texas	0	7*	0	-9*	0	9*	3	7	6	-1
Anchorange, Alaska	9*	5*	4	1	6*	2	12	5	11	5
Rockland, Maine	5	3*	-3	-2	-5*	-2	10	3	3	3
Las Vegas, Nevada	-13*	-6*	-14*	-5*	-13*	-5*	-14	-6	-7	-5
Fresno, California	2	-3*	-1	-4*	-1	-4*	1	-3	-1	-1
Philadelphia, Pennsylvania	-2	-6*	-2	-3*	-	-2	-3	-6	0	-2
Taft, Texas	-4*	1	-7*	-1	-7*	-2	-2	1	-3	[illegible]
Grand Rapids, Michigan	-2	-2	-3	-2	-2	-3*	1	-2	4	-2
Hartford, Connecticut	-6*	-1	-6*	-2	-6*	-2	-3	-1	2	[illegible]
McComb, Mississippi	-5	1	-5*	-1	-2	0	-2	+2	4	4
Seattle, Washington	-3	-3*	-6*	-6*	-6*	-5*	0	-3	10	0
Portland, Maine	-4*	-3*	-6*	-4*	-7*	-4*	-2	3	-3	-3
Jacksonville, Florida	5*	6*	5*	6*	4*	5*	4	6	6	5
Hammond, Indiana	-1	-1	-3	-2	-2	-1	1	-1	2	0
Bronx, New York	-11*	-3	-10*	-2	11*	-1	-9	-3	-2	4

See footnotes to Table VI-1.

Table VI-3

Difference in Achievement Test Raw Scores Between Experimental and Control Students by Site, Grade and Subject - Grade 3

	Mean Gain Difference[a]		Regression Model III[b]		Regression Model IV[c]		Adjusted Mean Gain Difference[d]		Adjusted Pooled Mean Gain Difference[e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer, Tennessee	11*	15*	10*	13*	9*	12*	12	17	8	12
Athens, Georgia	5*	5*	2	4*	3	4*	7	7	6	3
Wichita, Kansas	3	-6	-2	-6*	-2	-6*	6	6	0	-2
Dallas, Texas	6*	20*	7*	14*	7*	15*	5	20	-3	10
Anchorange, Alaska	5*	1	0	-1	0	0	1	4	-2	-5
Rockland, Maine	8*	3	1	2	0	-2	0	-1	3	3
Las Vegas, Nevada	-7*	-12*	-7*	-13*	-7*	-13*	-6	11	1	-1
Fresno, California	0	-5*	-2	-7*	-2	-6*	1	4	-2	-5
Philadelphia, Pennsylvania	1	-1	1	-2	2	-1	0	-1	3	3
Taft, Texas	1	7*	0	5*	0	5*	2	8	2	3
Grand Rapids, Michigan	-1	-2	-2	2	-6*	-8*	0	1	0	2
Hartford, Connecticut	4	-11*	0	-13*	0	-13*	8	-7	11	-7
McComb, Mississippi	-3	-8*	-8*	-11*	-6*	-9*	3	-2	5	4
Seattle, Washington	-6*	-3	-4*	-9*	-4	-9*	1	4	6	3
Portland, Maine	-1	-5*	-2	-7*	-2	-7*	1	-3	6	-2
Jacksonville, Florida	6*	7*	6*	7*	5*	6*	6	7	3	7
Hammond, Indiana	-2	-3	-6*	-5*	-6*	-5*	1	-1	1	-2
Bronx, New York	-5*	-5	-5	-4	-4	-4	-4	-6	6	7

See footnotes to Table VI-1.

Table VI-4

Difference in Achievement Test Raw Scores Between Experimental and Control Students by Site, Grade and Subject - Grade 7

	Mean Gain Difference[a]		Regression Model III[b]		Regression Model IV[c]		Adjusted Mean Gain Difference[d]		Adjusted Pooled Mean Gain Difference[e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer, Tennessee	1	2	3	2	2	1	0	C	1	0
Athens, Georgia	5*	5*	5*	4*	5*	5*	6	7	2	8
Wichita, Kansas	1	0	1	-	1	1	1	-1	0	-5
Dallas, Texas	2	-2	0	-1	0	-1				
Anchorange, Alaska	1	5	0	5*	0	-1	5	-3	-1	-4
Rockland, Maine	-1	-3	-6*	-6*	-6*	-6*	1	1	1	0
Las Vegas, Nevada	2	-4	3	-4*	2	-4*	2	-5	3	-1
Fresno, California	2	-9*	1	-7*	1	-7*	3	-8	3	-1
Philadelphia, Pennsylvania	-2	1	-2	1	-1	2	-2	0	2	3
Taft, Texas	0	0	-1	-2	-1	-3	1	3	1	-2
Grand Rapids, Michigan	1	-4	-1	-5*	-1	-4*	2	-3	2	-1
Hartford, Connecticut		0	1	0	1	0	1	1	2	3
McComb, Mississippi	2	0	2	1	2	0	2	0	2	4
Seattle, Washington	2	-12*	-1	-15*	0	-14*	3	-8	0	-7
Portland, Maine	2	2	1	1	1	2	3	4	2	3
Jacksonville, Florida	-2	-1	-3*	-3	-3*	-2	-2	0	0	1
Hammond, Indiana	1	2	-1	0	-1	0	0	4	3	1
Bronx, New York	4	5	6	5	6	5	4	5	4	-2

See footnotes to Table VI-1.

Table VI-5

Difference in Achievement Test Raw Scores Between
Experimental and Control Students by Site,
Grade and Subject - Grade 8

	Mean Gain Difference[a]		Regression Model III[b]		Regression Model IV[c]		Adjusted Mean Gain Difference[d]		Adjusted Pooled Mean Gain Difference[e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer, Tennessee	0	2	0	1	0	2	1	3	-2	-1
Athens, Georgia	-2*	-2	-3	-2	-2	-2	-1	-2	-2	-3
Wichita, Kansas	-1	1	-2	1	-2	1	0	3	0	2
Dallas, Texas	5*	7*	3	-1	1	-1	6	8	4	10
Anchorange, Alaska	8*	-3	7*	-4	6*	-4	10	-2	6	-1
Rockland, Maine	0	-2	-4*	-5*	-5*	-6*	3	2	2	1
Las Vegas, Nevada	2	-2	3	-1	2	-1	2	-2	-1	-3
Fresno, California	-3*	-3	-3*	-2	-4*	-2	-2	-1	-2	-1
Philadelphia, Pennsylvania	-5*	1	-8*	-1	-6*	1	-5	4	3	1
Taft, Texas	0	1	-1	0	-2	0	1	3	0	0
Grand Rapids, Michigan	7	-7	5*	-8*	6*	-4*	8	-6	3	-4
Hartford, Connecticut	-1	-2	-1	-2	1	-1	-2	-3	1	-4
McComb, Mississippi	-2	-3	-3	-4*	-2	-3	-1	-1	3	2
Seattle, Washington	-9*	-19*	-10*	-19*	-10*	-20*	-8	-18	-4	-10
Portland, Maine	-1	-6*	-2	-6*	-1	-5*	0	5	2	-3
Jacksonville, Florida	2	0	1	1	2	0	2	1	0	-1
Hammond, Indiana	-2	-3	-4*	-3*	-3*	-3	0	-1	1	0
Bronx, New York	-2	-3	-3	-2	-2	-2	-2	3	2	1

See footnotes to Table VI-1.

Table VI-6

Difference in Achievement Test Raw Scores Between
Experimental and Control Students by Site,
Grade and Subject - Grade 9

	Mean Gain Difference[a]		Regression Model III[b]		Regression Model IV[c]		Adjusted Mean Gain Difference[d]		Adjusted Pooled Mean Gain Difference[e]	
	Read	Math	Read	Math	Read	Math	Read	Math	Read	Math
Selmer, Tennessee	3	-3	0	-5*	1	-5*	4	-2	0	2
Athens, Georgia	3*	5*	2	4*	2	5*	3	6	3	7
Wichita, Kansas	0	-1	-1	-3	-1	-3	0	0	0	2
Dallas, Texas	0	1	-1	1	-1	1	0	1	0	1
Anchorage, Alaska	5*	11*	2	10*	1	9*	6	11	0	6
Rockland, Maine	5*	9*	0	3	0	2	6	13	3	2
Las Vegas, Nevada	6*	2	4	1	3	0	6	2	3	2
Fresno, California	4	2	2	1	2	1	4	3	4	-1
Philadelphia,Pennsylvania	-4*	-3	-4*	-3	-4	-3	-4	-3	-4	-2
Taft, Texas	8*	4*	5*	7*	6*	7*	8	3	2	0
Grand Rapids, Michigan	4	-1	3	-1	1	-2	4	-1	2	-1
Hartford, Connecticut	1	4*	1	3	1	1	1	5	3	2
McComb, Mississippi	0	-1	-2	-3	-2	-3	1	1	2	1
Seattle, Washington	3	2*	4*	10*	-3	-9*	-3	-12	6	-2
Portland, Maine	-1	-5*	-2	-5*	-1	-4*	-1	-5	1	0
Jacksonville, Florida	2	1	2	1	3	1	2	1	3	1
Hammond, Indiana	1	0	-1	-2	-1	-1	2	1	1	5
Bronx, New York	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

See footnotes to Table VI-1.

last two columns present the mean gain differences between the experimental group at a particular site and the control group from all sites combined, adjusted by expression (19). A comparison of the entries in these columns with those in the ninth and tenth columns enables us to identify the cases where large positive or negative experimental-control differences may be a result of abnormally small or large gains on the part of the local control group or site measurement error for the control students. We stress the word "may" because a gain score at a particular site which appears to be abnormally large or small in relation to the average gain score across all sites may not be abnormally large or small for that particular site. Since we do not have the data to ascertain the degree to which "normal" gains vary from site to site, cases where there are differences between the pooled and unpooled adjusted mean gain differences are difficult to interpret. The existence of such differences is still another reason why we have less confidence in the site-by-site results.

Note first of all that differences between experimental and control mean gains in columns 1 and 2 are much larger at particular sites than they are in the overall results. Differences of 5 or more raw score points are quite common, particularly in the elementary grades, and there are 24 cases of differences of 10 or more raw score points.

Once again the coefficients in the two regression models are nearly identical in all of the site-grade-subject combinations. 13/ These regression coefficients are normally but not always more negative than

---

13/ These results are roughly consistent with the site-by-site results reported by Battelle (2).

the mean gain differences. In some cases like the first grade in Jacksonville, the experimental group mean pretest scores exceeded that of the control group instead of the more normal reverse case. Unlike the aggregate results, however, experimental-control differences of 5 or more raw score points are common. And there are 28 cases of differences of 10 or more raw score points. Also, unlike the aggregate results,in many site-grade-subject cases the difference between the mean gain comparisons and the regression coefficients is quite large. The larger differences in the site by site results correspond to the larger differences between pretest group means at the site level.

As in the aggregate results, both the mean gain differences and the regression coefficients are biased. Therefore, we have again adjusted the simple mean gain differences on the basis of our best overall estimate of α_1. 14/ We note that while the adjusted mean gain differences presented in columns 9 and 10 generally reduce the experimental-control differences by up to a few raw score points, they are on the whole quite similar to the unadjusted differences.

The pooled adjusted differences presented in columns 11 and 12, however, sometimes differ substantially from the adjusted differences. Out of a total of 76 adjusted mean gain differences of 5 raw score points or more, 23 have dramatically different pooled adjusted mean gain

---

14/ We could have tried to estimate α_1 separately for each site. Since the sample size at each site is approximately 1/18 the size of the total sample, however, it is not clear that this alternative procedure would have led to more reliable adjusted estimates.

differences. The relatively large changes occur in Anchorage math, Philadelphia reading, Hartford math, Seattle reading, Jacksonville reading and math, and Hammond reading in the first grade; Dallas math and Bronx reading in the second grade; Dallas reading, Las Vegas reading and math, and Bronx math in the third grade; Dallas reading, Anchorage math, Fresno math, and Bronx math in the seventh grade; Philadelphia reading the eighth grade; and Anchorage reading, Rockland math, Taft reading, Seattle math, and Portland math in the ninth grade. Thus almost one-third of the apparent relative failures and successes are open to question. There are nine additional cases where the pooled adjusted mean gain differences are equal to 5 raw score points or more, while the corresponding adjusted mean gain differences were not nearly so large. These relatively large changes occur in Fresno reading and Taft reading and math in the first grade; Dallas reading and Seattle reading in the second grade; Seattle and Portland reading and Bronx reading in the third grade; and Seattle reading in the ninth grade. Thus abnormally large or small control gains may also be obscuring a few cases of relatively good or bad performances of a contractor.

In order to evaluate the educational significance of these raw score differences, in Table VII we have translated the adjusted mean gain differences into the corresponding differences in grade equivalents. (The differences are evaluated at the experimental post-test mean score.) Cases where significant differences in the adjusted mean gain differences are eliminated by pooling control students are denoted by a single

TABLE VII

Difference in Adjusted Mean Gain Grade Equivalents Between Experimental and Control Students by Site, Grade and Subject[a]

	GRADE 1		GRADE 2		GRADE 3		GRADE 7		GRADE 8		GRADE 9		AVERAGE
	R	M	R	M	R	M	R	M	R	M	R	M	
Selmer, Tennessee	.9	.2	0	.1	.5	.5	0	0	.1	.4	.6	-.2	.26
Athens, Georgia	0	.2	0	0	.2	.3	.4	.5	-.1	-.2	.4	.6	.19
Wichita	-.8	-.1	.2	.1	.2	.1	0	0	0	.1	0	0	-.02
Dallas, Texas	.4	.4	.1**	.1*	.1*	.6	.5*	-.2	.7	.9	0	0	.30
Anchorage, Alaska	.1	.4*	.2	.3	.2	.2	0	.4*	1.1	-.2	.5*	1.0	.35
Rockland, Maine	-.1	0	.2	.2	.4	.4	0	.1	.4	.1	.7	1.1*	.30
Las Vegas, Nevada	0	-.9	-.3	-.2	-.2*	-.5*	.2	-.3	.2	-.2	.7	.2	-.10
Fresno, California	0	-.7	0	-.2	0	-.1	.4	-.4*	-.2	-.1	.6	.4	-.01
Philadelphia, Pennsylvania	-.7*	-.9	-.1	-.3	0	0	-.2	0	-.7*	.4	-.5	-.4	-.28
Taft, Texas	0	-.1	0	.1	0	.4	.1	.1	.1	.3	1.1*	.2	.18
Grand Rapids, Michigan	.3	-.3	.1	-.1	0	-.1	.1	-.2	.9	-.6	.5	0	.06
Hartford, Connecticut	-.2	-.6*	0	0	.2	-.1	0	0	-.2	-.2	.1	.5	.06
McComb, Mississippi	.4	1.0	0	0	.1	-.1	.2	0	-.1	-.2	.2	0	.12
Seattle, Washington	-.8*	-.3	0	-.2	.1**	0	.1	-.5	-.9	-1.4	-.1**	.9*	-.41
Portland, Maine	-.4	-.6	0	-.2	0**	-.1	.3	.3	0	-.2	-.2	-.5*	-.13
Jacksonville, Florida	.7*	.7*	0	.4	.2	.2	-.2	0	.2	0	.3	0	.21
Hammond, Indiana	-.5*	0	0	0	0	0	0	.3	0	0	.2	.1	.01
Bronx, New York	NA	NA	.2*	-.2	-.1**	-.2*	.2	.2*	-.3	-.4	NA	NA	-.08
Average	.03	-.09	.03	0	.10	.10	.12	.02	.07	-.08	.30	.12	.05

[a] Derived from Columns 7 and 8 in Table VI. The mean gain differences in grade equivalents are evaluated at the experimental mean post-test raw score.

*Indicates that significant differences in adjusted mean gain differences are eliminated by pooling control students

**Indicates that insignificant differences in adjusted mean gain differences become significant when control students are pooled.

asterisk, while cases where large differences only emerge for the pooled adjusted mean gain differences are denoted by two asterisks. As a crude method of summarizing the results, we present in the last row average differences in gains across sites for each grade and subject and in the last column average differences in gains across grade and subject for each site. The averages across sites for each grade-subject case are merely another way of presenting our aggregate results of Table III--though they are not perfectly consistent because of nonlinearities in the translation of raw scores to grade equivalents. But the averages for each site in the last column are quite interesting.

A cursory look at these averages suggests that the overall mild effect is being produced by the offsetting of some relatively successful sites by some relatively unsuccessful sites. Furthermore, some of the average relative gains and losses appear to be noteworthy, including a forty percent of one grade equivalent loss in Seattle and 35 percent of one grade equivalent gain in Anchorage. However, a more careful examination of the Table suggests that some of the largest apparent winners or losers may be artificially inflated because of either control student volatility or control measurement error problems. We note especially the fact that the sites with the largest average differences in gains also tend to have the greatest number of asterisks. If we had used the pooled adjusted mean gain differences rather than the adjusted mean gain differences, the largest average differences in gains would have been appreciably smaller. Thus while it seems that performance contracting did work somewhat better than the normal public schools at some sites and somewhat worse at others, the magnitude of these relative successes and failures was generally small, and even if large, not fully trustworthy.

While there are some regional, city size, and company patterns to the results, at this point we can do no more than note them and suggest caution in their interpretation. Southern cities and small cities generally fared much better than large and non-Southern cities. Two companies--in the first two sets of three sites in Table VII--appear to have done somewhat better than the normal public schools; two companies--in the fourth and sixth sets of sites--appear to have done just barely better than the public schools, and two other companies -- in the third and fifth set of sites--appear to have done somewhat worse than the public schools. But it is important to note that since different companies did not run programs in the same sites, it is conceptually impossible to disentangle the site from the company effects.

V. Conclusion

Despite the problems that inevitably accompany anything as complicated as a large-scale experiment in performance contracting--the difficulties of testing human beings, the imperfect matching of experimental and control students, and other uncertainties--these results are remarkably consistent. Our analysis almost always indicates that there were no signifcant differences in the achievement gains of the experimental and control groups. Not only did both groups do equally poorly in terms of overall averages, but also these averages were very nearly the same in each grade, in each subject, and for the best and worst students in the sample. There were some successes and failures among the individual sites, at least in certain grades and subjects, but even many of these are statistically quite unreliable--possibly caused by the volatility of control students or site-wide testing difficulties. Indeed, probably the most interesting aspect of these conclusions is their very consistency. This evidence indicates with surprising uniformity that the performance contractors who participated in the experiment do not currently have the capability of bringing about any great improvement in the educational status of disadvantaged children.

Appendix I

Measurement Bias in Regression Coefficients when Other Variables are Included

Assume we have the model

(1) $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3^*$,

where X_3^* is the imperfectly measured variable, or

(2) $X_3 = X_3^* + w$,

and X_2 is the other variable we have added.

The true coefficients of (1) are given by

(3) $\beta_1 = \bar{Y} - (\beta_2 \bar{X}_2 - \beta_3 \bar{X}_3)$

$$\beta_2 = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \quad \frac{VAR(Y)}{VAR(X_2)}$$

$$\beta_3 = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \quad \frac{VAR(Y)}{VAR(X_3^*)}$$

But if we try to estimate (1) using observed values of X_3, we know that since $p = VAR(X_3^*)/VAR(X_3) \ll 1$, we will get estimates of r_{13} and r_{23} which will be biased towards zero ($\hat{r}_{13} < r_{13}$; $\hat{r}_{23} < r_{23}$). The corresponding expressions for the coefficients when estimated with measured data are then

(4) $$b_1 = \bar{Y} - \left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left[\left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right)\beta_2 \bar{X}_2 + \left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right) p\,\beta_3 \bar{X}_3\right]$$

$$b_2 = \left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{1-\hat{r}_{23}^2}\right)\frac{VAR(Y)}{VAR(X_2)} = \left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right)\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\beta_2 \gtrless \beta_2$$

$$b_3 = \frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{1-r_{23}^2} \quad \frac{VAR(Y)}{VAR(X_3)} = \left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right)\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right) p\,\beta_3 \gtrless p\,\beta_3.$$

Substituting (3) into (4) gives

$$(5)\quad b_1 = \beta_1 + \beta_2\bar{X}_2 \left[1-\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}}\right)\left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right)\right]$$

$$+\ \beta_3\bar{X}_3 \left[1-p\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right)\right].$$

The estimated experimental effect is then derived by substracting b_1^C from b_1^E, or

$$(6)\quad b_1^E - b_1^C = \beta_1^E - \beta_1^C + \beta_2\left(\bar{X}_2^E - \bar{X}_2^C\right)\left[1-\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right)\right]$$

$$+\beta_3\left(\bar{X}_3^E - \bar{X}_3^C\right)\left[1-p\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right)\right].$$

We remember from the paper that the bias when $\beta_2=0$ was $\beta_3\left(\bar{X}_3^E - \bar{X}_3^C\right)(1-p)$. If $B_2 > 0$ and $\bar{X}_2^C > \bar{X}_2^E$, therefore, we must have some combination of

$$(7)\quad \left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right) > 1 \text{ or } b_3 > p\,\beta_3$$

or

$$\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{r_{13}-\hat{r}_{12}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right) > 1$$

to reduce this bias. To eliminate the bias altogether we must have

$$(8)\quad \frac{\beta_2\left(\bar{X}_2^E - \bar{X}_2^C\right)}{\beta_3\left(\bar{X}_3^E - \bar{X}_3^C\right)} = -\frac{\left[1-p\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{\hat{r}_{13}-r_{12}\hat{r}_{23}}{r_{13}-r_{12}r_{23}}\right)\right]}{\left[1-\left(\frac{1-r_{23}^2}{1-\hat{r}_{23}^2}\right)\left(\frac{r_{12}-\hat{r}_{13}\hat{r}_{23}}{r_{12}-r_{13}r_{23}}\right)\right]}.$$

These are very stringent conditions and it is not at all obvious that they will be satisfied.

References

1. Bartlett, M.S., "The Fitting of Straight Lines If Both Variables Are Subject to Error," Biometrics, Vol. 5, 1949.

2. Battelle Columbus Laboratories, Office of Economic Opportunity Experiment in Educational Performance Contracting, Jan., 1972.

3. Carpenter, Polly and George R. Hall, "Case Studies in Educational Performance Contracting: Conclusions and Implications," Rand Corporation, December, 1971.

4. Goldberger, Arthur S., Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations, Discussion Paper, Institute for Research on Poverty, University of Wisconsin, 1972.

5. Johnston, J., Econometric Methods, McGraw-Hill, 1963.

6. Office of Economic Opportunity, An Experiment in Performance Contracting: Summary of Preliminary Results, February, 1972.

7. Saupe, J. L. "Some Useful Estimates of the Kuder-Richardson Formula Number 20 Reliability Coefficient," Educational Psychological Measurement, Vol. 21, 1969.

8. Wald, A.,"The Fitting of Straight Lines If Both Variables Are Subject to Error" Annual of Mathematical Statistics, Vol. 11, 1940.

- 51 -

Chapter II

IMPLICATIONS OF USING STANDARDIZED TESTS IN PERFORMANCE CONTRACTING

by

Jeffry S. Schiller

and

Ellen Press Murdoch

INTRODUCTION

Standardized Achievement Tests and Performance Contracting

School systems throughout the nation administer standardized achievement tests to millions of children to assess how much they have learned as a result of their school experiences. In recent years, controversy has developed concerning the uses and abuses of such measures. Discussion has centered around such questions as:

1. Do these tests in fact measure what the students are taught? Indeed, can this be measured?
2. Are these tests reliable enough for the purpose of assessing student performance? Can individual performance be reliably measured? Group performance? Can standardized test scores be utilized for measuring change over a period of time? If so, for groups and/or individual students?
3. What technique of interpreting test scores is best? Does it make any difference if test scores are reported in percentiles, grade level equivalents, stanines, or raw scores?

While questions such as these have been debated for many years, the use of standardized tests in performance incentive contracting experiments has intensified the debate. Standardized tests were used for two purposes in the OEO experiment. An evaluation test was administered to both experimental and control groups in the fall and spring.

The results of this test were used to assess the overall impact of the performance contracting program. A second set of standardized tests administered only to the experimental groups was used to determine the firms' payments. Seventy-five percent of the payment to contractors was based on how much individual students improved on these tests from fall to spring. In addition, criterion referenced tests developed by the contractors were administered at five times during the program year. Twenty-five percent of the payment to contractors was based on student performance on these criterion referenced tests.

The basic issue with respect to the use of standardized tests in performance incentive contracting has been:

Are standardized achievement tests sufficiently precise instruments to allow for the assessment of an individual student's progress over time for purposes of (a) assessing the impacts of performance contracting programs? and (b) computing the number of dollars to be paid a contractor for that student?

To understand the issues that have arisen from the use of standardized tests in the experiment, it is necessary to review some key concepts, including:

1. Method of standardized achievement test construction
2. Scoring techniques
3. Criteria for test selection
    a. Validity
    b. Reliability and standard error of measurement

STANDARDIZED TEST CONSTRUCTION AND SCORING

Test Construction

The first step in the construction of standardized achievement tests is deciding on the subject areas that the test will cover. Since the authors of achievement tests generally want to measure what students have been taught, they begin by determining what is generally taught in a given subject in a particular grade throughout the country. After studying curriculum guides, textbooks, and statements of objectives from various school systems, as well as consulting with specialists in the subject areas, the authors develop a test outline which specifies the concepts to be covered and the amount of emphasis to be given to particular aspects of the material. The items (questions) for the test are then written, with each item designed to test a student's knowledge of some aspect of the material. These items are then reviewed, edited, and assembled into a preliminary form of the test.

Once the preliminary form is ready, the authors undertake an "item analysis program" to determine if the items they have written are "good" items and of an appropriate level of difficulty for the group for which the test is intended. The preliminary form is administered to a group of students selected to be representative of the students who will use the final form of the test. In addition, the preliminary form is usually administered to students one grade level above and one grade level below those for whom the final form is intended.

The preliminary forms are then scored, and the percentage of students in each grade answering each item correctly is computed. In general, most of the items for a third grade test will be items which 40 to 60 percent of the students in third grade in the item analysis program answered correctly. Some more difficult and some less difficult items are also included. The authors of these tests want the final form of the tests to be one on which "good" students will receive higher scores than "poor" students and one on which fourth graders will receive higher scores than third graders. Items which do not "discriminate" are therefore eliminated. That is, if more "poor" students answer an item correctly than "good" students, or if more third graders answer an item correctly than fourth graders, the item is eliminated. A "good" student, for the purposes of the item analysis program is one who receives a high overall score on the preliminary test.

A test for which a thorough item analysis has been done should thus contain items of an appropriate level of difficulty which discriminate between good and poor students at a given grade level and which test the student's knowledge of the material contained in the outline of the test. Once the final form of the test is ready and directions for administering and scoring the test have been prepared, the test is "normed." In the case of a nationally normed test, the test publisher wants to provide information which will enable the test user to compare the performance of his students with the performance of students nationally.

For this reason, the publisher attempts to select a sample of students which is representative of all students in a particular grade throughout the country. Since it is practically and operationally impossible to select a group which is representative of all students in the country in all respects, the norm samples usually include students who are representative with respect to several characteristics assumed to be related to school achievement and for which information is readily available. Attempts generally are made to include students from various geographic regions, communities of different socio-economic status, and school systems of varying size in proportion to their numbers in the national population. Publishers also attempt to select a norm sample which is representative with respect to IQ scores.

The final form (s) of the test are then administered to the students in the norm sample, and scored. The various normative score tables provided by the publisher with the test are based on the score distributions from this administration of the test.

Standardized Test Scoring:

The most direct way of describing a particular student's performance on a test is in terms of his raw score, or how many questions he answered correctly. A raw score in itself does not provide information on "how well" the student did compared to others and does not provide school personnel with a meaningful frame of reference. The significance of a raw score will vary, depending on how difficult the test was or how many items were on the test. Such a score does become meaningful when it is compared to scores of other students taking the test.

One method of interpreting a raw score is in terms of its corresponding standard score. A standard score expresses a particular student's performance in terms of how many standard deviation units his score is above or below the means of the test. The standard deviation is a measure of the variability in a distribution of scores and is expressed in test score units. A distribution in which the scores are clustered close to the mean will have a smaller standard deviation than a distribution where many scores vary a great deal from the mean of the test.

In a normal distribution, one would expect the scores of approximately two-thirds of the students to fall within one standard deviation on either side of the mean.

Stanine scores are a particular type of standard score. The score scale is divided into nine bands with each band including scores within one-half of a standard deviation (except for the 1st and 9th stanine). The middle band, or fifth stanine, includes scores from one-fourth of a standard deviation below to one-fourth of a standard deviation above the mean. The fourth stanine includes scores from one-fourth to three-fourths of a standard deviation below the mean while the sixth stanine includes scores from one-fourth to three-fourths of a standard deviation above the mean, and so on, with the first stanine including extremely low scores and the ninth stanine including extremely high scores.

Identifying a student as being in a particular stanine therefore expresses his position relative to other students in terms of standard deviation units above or below the mean of the test.

Another way of comparing a student's performance to other students is to assign a percentile rank to particular raw score values. On the basis of the score distributions of the norm sample, the publisher determines, for each raw score value, the percentage of students with raw scores equal to or lower than that particular value. For instance, if a raw score of 26 corresponds to a percentile rank of 43, it means that 43 percent of the students in that grade received a raw score of 26 or less. While percentile ranks provide information about a student's performance relative to the performance of other students in his grade, they are not generally suitable for measuring student progress because the gain or loss of 3 percentile ranks means different things at different points along the scale. This distortion occurs because in a normal distribution, more scores occur close to the mean than at either the upper or lower end of the distribution. Therefore, a gain of one raw score point means something different in terms of percentiles if the student is near the mean of the distribution than if he is at one of the extremes. For instance, the difference between the number of questions answered correctly at the 50th percentile and at the 55th percentile may be very small while the number of questions answered correctly between the 5th percentile and the 10th percentile may be very great.

The most commonly used and easily understood method of interpreting a student's score is the grade equivalent. These scores characterize a student's raw score as equivalent to the median score of students at a particular grade level. They are obtained by administering one test to several successive grades and determining what the median scores are for the various grades in the norm sample. For instance, if a test is normed on fifth graders in the second month of school and the median score in raw score units is 56, a raw score of 56 corresponds to a grade equivalent of 5.2: fifth grade, second month. If a test is normed in the third month of the school year, the median score of the fifth graders in the norm sample corresponds to a grade equivalent of 5.3, the median score of the sixth graders to a grade equivalent of 6.3 and so on. The raw score values corresponding to 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 6.0, 6.1 and 6.2 are assigned by distributing raw score values between the value obtained for 5.3 and the value obtained for 6.3 over the remaining grade equivalent values. In dealing with grade equivalent scores, it is important to keep in mind two points. First, in developing these grade equivalent scales, test-makers divide a school year into 10 units and assume that learning progresses evenly throughout the year. That is, they assume that a student learns the same amount in the third month of fifth grade as he learns in the eighth month. Second, the fact that a fifth grade student receives a grade equivalent score of 8.0 on a test intended for fifth graders does not mean that the student has learned everything there is to learn in fifth, sixth, and seventh grades and is therefore ready to begin eighth grade. If the same student were given a test which was intended for use by eighth graders and which covered content appropriate for eighth graders, it is possible that his grade equivalent score would be substantially lower than 8.0. In developing grade equivalent scales, by extrapolation or the use of overlapping tests or linking tests publishers often will develop a grade equivalent scale which includes grade equivalents for grades above and below those actually included in the norming program. The extent to which a grade equivalent scale is based on actual administration of the test to the grades for which grade equivalents are provided is an important factor in determining the "validity" of a particular grade equivalent scale.

Any discussion of the use of grade equivalents in an experiment such as performance contracting must focus on the number of raw score points needed to raise a student's performance by one grade level. (Most of the contracts in this experiment stipulated that the contractors would be paid only for students who made grade equivalent gains of 1.0 or more.) It has been pointed out that in some cases a child need answer only two, three, or four more questions correctly on the post-test than he did on the pre-test to gain a full grade level. For the most part this occurs at either the very high and/or very low end of the raw score distribution.

One explanation for this has to do with the way in which the tests are constructed. In developing a test for third graders, the publisher is most interested in including items which will discriminate between second and third graders and third and fourth graders since he expects that these items will provide maximum information about most of the students who will take the test. In the interest of making the final test a suitable length, he will probably include a great many more items which distinguish between second and third and third and fourth graders than items which distinguish between seventh and eighth graders. If he administers the final form of his test to students in grades three through eight, it is not, then, surprising to find a very small difference between the average score of the seventh graders and the average score of the eighth graders in raw score points. The same is true for the lowest grades.

Following are examples of the relationship between raw scores and grade equivalents that occurred in one test used in the OEO experiment. These data are very similar to those from the other tests used in the experiment and illustrate the nature of the raw score to grade equivalent relationship (Appendix A displays data for all of the tests used in the experiment.)

Example #1: Metropolitan Achievement Test, Primary I, Reading, 1970 edition, used with 2nd graders.

a. To increase from 1.0 to 2.0 in terms of grade equivalents, a student's raw score must increase from 10 to 51 points (41 additional questions.)

b. If a student's raw score increases from 11 to 16 points, it makes no difference in grade equivalents at all. Raw scores from 11 to 16 all yield a grade equivalent of 1.1. In this case, the contractor would not receive any payment for an improvement of 5 raw score points.

c. However, for a student to progress from a grade equivalent of 3.1 to 5.0, he need only answer three additional questions. (An increase in raw score points from 74 to 77.)

Example #2: Metropolitan Achievement Test, Intermediate Battery, Reading, 1970 edition, used with 7th graders.

a. To increase from 3.0 to 4.0 grade equivalents, a student must answer 11 additional questions correctly (an increase from 23 to 34 in raw score points); from 4.0 to 5.0, 12 questions (34 to 46 in raw score points); from 5.0 to 6.0, 11 questions (47 to 58 in raw score points); from 6.0 to 7.0 11 questions (59 to 70 in raw score points); from 7.0 to 8.0, six questions (71 to 77 in raw score points).

b. However, to increase from 8.0 to 9.0, only four additional questions need be answered correctly (77 to 81 in raw score points) and only three questions to increase from 9.0 to 10.0 (81 to 84 in raw score points).

Because a very small increase in raw score points can result in a large grade equivalent gain at certain points on the scale, it has been pointed out that a contractor paid on the basis of grade equivalents gains might be rewarded for very little improvement in terms of raw score points. It is important to bear in mind, however, that where the scores of a particular group fall on the scale on the pre-test will determine the degree to which the raw-score-point grade-equivalent relationship works in favor of the contractor. Typically, students with scores at or near the mean of the standardization sample on the pre-test will have to show considerable improvement in terms of raw score points to make a grade equivalent gain of 1.0. Because the distribution of a subject population's pre-test scores contributes to the extent to which small raw score gains will result in large grade equivalent gains, the selection of an appropriate test level becomes extremely important.

Despite their imperfections, the advantages of using grade equivalents are numerous. They are easily understood by the public and by school personnel. They also provide both baseline and gain score information. That is, even though the grade gain concept has some distortion, its use on a year-to-year basis by school personnel provides a baseline of previous information with which to compare current results.

In the OEO experiment, grade equivalent gains were used to compute payments to contractors. For evaluation purposes, raw score differences between experiment and control groups were used, although the results are presented in terms of grade equivalents.

CONSIDERATIONS IN STANDARDIZED TEST SELECTION

In deciding whether a standardized test is appropriate for a particular purpose, there are basically two questions which must be addressed.

- -- Is it valid? Does it measure what we intend to measure?
- -- Is it reliable? Does the test consistently measure the characteristics it purports to measure?

Whether a particular test is valid depends on what it is being used to measure. In this experiment, we wanted to assess how well the students learned what is generally taught in the areas of reading and mathematics. While the contractors were basically free to decide how they would attain certain objectives, the decision as to what the objectives would be was not theirs to make. Their agreement to allow OEO to use multiple standardized achievement tests and to eliminate the identification of these tests was indication of their belief that in general, standardized achievement tests are a fair and adequate measure of what they were teaching in reading and mathematics and that there is a great deal of overlap in the content of various test batteries. The contractors were asked to indicate in their proposals several standardized achievement tests which they would recommend for measuring the impact of their program. Almost every test used was recommended by one or more contractors.

Since the tests are, for the most part, wide surveys of curriculum and standardized on national samples, we are confident that they provide the best measure of "what is generally taught." The fact that success on these tests appears to be related to success in school was also important in that we felt by using these tests we were setting worthwhile and fairly broad objectives for the contractors.

Reliability is concerned with the degree to which a test consistently measures whatever it is in fact measuring. It addresses the question of: Are the test scores stable? Are the scores the students obtained on the test an adequate indication of their true scores on the test?

While there is no single way of estimating test reliability, the following four techniques are generally used in obtaining reliability coefficients for standardized tests:

1. Administering a test to a group of students, retesting them with the same test after a brief interval, and computing a correlation between the two tests.
2. Administering two different forms of the same test to a single group of students and computing a correlation between them.
3. Dividing a single test into two equal parts, administering the test to a group of students, and computing a correlation between the two halves of the test.
4. Examining the consistency of response from item to item on a single test. Formulas such as the Kuder-Richardson # 20

and #21 are used in exmining the internal consistency of tests.[1]

The estimate of the reliability of a test can vary depending on the technique used and the type of group used in computing it. For this reason, it is important to look not only at the estimate itself, but also at the sample and method used in obtaining it, especially if one is comparing several different tests to determine which is the most reliable for use with a particular group.

---

[1] The KR20 formula yields a coefficient which can be expected to equal the mean of all possible split-half coefficients obtainable for a test. The formula is:

$$r = \frac{n}{n-1} \quad x \quad \frac{s_t^2 - \sum pq}{s_t^2}$$

where:
r = reliability
n = number of items
s_t = standard deviation of total test scores
q = proportion of students failing each item
p = proportion of students answering each item correctly

The KR_{21} formula is used with tests where all items are designed to measure a single ability and are of equal difficulty. If there is variation in item difficulty and a KR_{21} reliability estimate is used, however, the reliability of the test will be underestimated. The KR_{21} formula is:

$$r = \frac{n}{n-1} \quad x \quad \frac{s_t^2 - n\overline{pq}}{s_t^2}$$

where:
n = number of items
s_t = standard deviation of total test scores
$\overline{p} = \frac{\text{arithmetic mean of test scores}}{n}$
$\overline{q} = 1 - \overline{p}$

While there is no absolute standard for how reliable a test must be for particular purposes, there are some opinions on the subject. Nunnally (1967)[2] states that, when one is dealing with group scores, increasing reliability beyond .80 is often wasteful for basic research purposes. When important decisions are to be made on the basis of individual scores, he states that .90 is the minimum acceptable reliability and that a reliability of .95 is desirable.[3]

Test users very often need to know to what extent a score obtained on a given test is a dependable estimate of what a particular child can do on a test. Because the reliability coefficient in itself does not directly assist us in assessing that, the standard error or measurement (SEM), a statistic related to the reliability coefficient, is used.[4]

---

[2] Jum C. Nunnally, Psychometric Theory, New York, McGraw-Hill, 1967, p. 226

[3] According to Nunnally, "The alternate form method of measuring reliability is the ideal because it measures more sources of reliability and measures them better than any other method which is used. If it were not for practical difficulties, the alternate form method would be used in most instances." (Jum C. Nunnally, Basic Principles of Measurement and Evaluation, p. 84)

[4] The typical formula for computing the standard error of measurement for a test is $\text{SEM} = \text{S.D.}\sqrt{1 - \text{reliability coefficient}}$ where S.D. is the standard deviation of the scores on the test.

Because of the relative imprecision of test scores, it is generally agreed that individual scores should be interpreted as regions or bands, rather than points. By taking students' estimated "true" scores and marking off one SEM above and below the estimated "true"scores, we can establish a band and expect that in two-thirds of the cases, the student's "true" score will be somewhere in that band of scores.

A student's estimated true score is determined in the following way:

Estimated true score =
  reliability coefficient x obtained score,
  where scores are expressed in terms of
  deviations from the mean

For example, to estimate the "true" score of a student with a raw score of 70 on a test with a mean of 80 and a reliability of .90, we would begin by expressing his raw score in terms of deviation units from the mean. In this case, his score in deviation units would be -10, since his raw score of 70 is 10 points below the mean of the test (80). His estimated true score is .90 (reliability coefficient) x -10 (obtained score expressed in deviation units), or -9 (in deviation units). His score of -9 in deviation units is equal to 71 in raw score points, because it is 9 points below the mean of the test. If the SEM for this test is 3 (raw score points) we would mark off a band from 3 points below his estimated true score to 3 points above his estimated true score, or from 68 to 74. In two out of three cases, we would expect that the student's true score is between 68 and 74.

It should be noted that an interval established using the estimated true score is different than an interval established using the obtained score.

The use of the estimated true score in establishing the interval corrects for the fact that high scores tend to be biased upward and low scores tend to be biased downward. While the difference between the two intervals is not great in this example, it can become appreciable if a test has lower reliability and the obtained score is further from the mean.

The SEM and estimated "true" scores, then, can be used to establish a band of scores where one would expect a student's "true" score to fall. The use of the SEM helps to inject caution into the interpretation of small differences between raw scores.

It should be mentioned that the reliability of gain scores is lower than the reliability of either the pre- or post test. In addition, the SEM of a gain score will generally exceed the SEM of either test.[5]

The issue of gain score reliability is intensified in a performance contracting experiment because contractors are paid on the basis of these gain scores, and measurement errors might unfairly penalize or reward contractors.

---

[5] Georgia Sachs Adams, Measurement and Evaluation, (Holt, Rinehart and Wilson), p. 94, July, 1966

Because the mean gain score of a group is considered to be more accurate than an individual student's gain score, it is often suggested that contractors be paid on the basis of group gain (i.e. a specified amount of money for each .1 grade equivalent of gain multiplied by the number of students). Paying contractors a specified amount for each month of gain and penalizing him the same amount for each month of loss on an individual student basis yields the same result as paying on the basis of group gain.

We initially believed that payment based either on group gain or graduated payments *and penalties was* not satisfactory in that it might encourage contractors to pay greater attention to fast-learners to the detriment of the slower learning students. For this reason, we decided to base payment on individual gain with contractors receiving no reimbursement for a student who did not achieve an established minimum gain, generally 1.0 grade equivalents. This minimum gain, or guarantee, was established to reduce the possibility of contractors being reimbursed for gains which were solely the result of measurement error.

A typical contract in the experiment was one for which the contractor was paid nothing if a student gained less than 1.0 grade equivalents and $75.00 if the student gained 1.0 grade equivalent. If a student gained more than 1.0 grade equivalents, the contractor was paid $75.00 plus $5.36 for each .1 grade equivalent above 1.0. (For example, for a student who gained 1.3 grade equivalents, the contractor was paid $75.00 + [3 x $5.36], or $91.08).

If we assume that the mean gain of a group is the best indication of "true" gain, any method of payment which resulted in either a higher or lower profit for a contractor than he would receive on the basis of group gain could be considered as unfairly rewarding or penalizing the contractor. As previously noted the decision was made to (a) pay contractors on the basis of an individual student's progress in order to assure that attention be paid to each student, and (b) to impose a level of gain below (or guarantee) which the contractor would receive no payment. Our initial impression was that this payment computation procedure would create a situation in which payments would reflect performance at least as adequately as if they were computed on a group basis and also would encouarge individualized instruction. But further consideration suggests that this may not have been the case and that payment on an individual basis differed from what payment on a group gain basis would have been. When individual gain scores are used for payment purposes they are, of course, subject to measurement error. Assuming that this error is unbiased, a student's observed gain score might be more or less than his true gain score. Some of the scores which are increased as a result of measurement error may actually be elevated to or above the minimum guarantee and, thus, result in payment to the contractor which he would not otherwise have received. Of course, some students who scored above the guarantee could also be expected to fall below and not qualify the contractor for payment.

The number of students moving above or below the guarantee level because of error is related to the size of the guarantee and to the distribution of scores for the entire subject population. In general, if the true mean gain score for the population is the same as the guarantee level, overpayment or underpayment for individual students due to measurement errors should balance out. The more disparate the mean gain score and the guarantee level, the more contractors may either be helped or hurt by the payment system used in the OEO experiment.

Since the experiment's subject population had a mean gain score significantly lower than the guarantee level, it is probable that more students moved above the guarantee level as a result of measurement errors than fell below it. If a large proportion of the distribution was significantly above the guarantee, the contractor would probably lose money for students for whom he would have been paid if there were no measurement error.

The next chapter, "Contractual Procedures" by Charles B. Stalford discusses in some detail the implication of the payment scheme employed in the experiment. Let it suffice that the payment system used does not assume that dollars are paid for only real gains. Measurement errors can either inflate or reduce payment to a contractors, depending on the size of the guarantee and the amount of the gain. In the writing of performance contracts, it would be important to attempt to estimate student gain and set a guarantee level when the effect of error on payments would tend to balance out. (Specifically, where payment on an individual basis would equal payment on the basis of mean group gain.)

It might be that payment based on group scores, more free of measurement error, is more fair than payment based on individual scores and would outweigh the need to use the payment process as a means of insuring individualized instruction.

Presumably, individualized instruction could be encouraged together with the use of group scores through some other type of contract language which prescribed a penalty if the variation in the score exceeded mutually acceptable limits.

SELECTION OF TESTS FOR THE EXPERIMENT

The criteria for selecting standardized tests used in the experiment were:

1. The norms for the test had to be based on a relatively recent sample having a reasonably large number of students representative of the national population.
2. The tests had to measure what is generally taught in the areas of reading and mathematics in school throughout the country based on a fairly recent survey of "what is taught."
3. The tests had to display a high degree of reliability.
4. The tests had to have very clear and simple directions for administration.

In addition to evaluating the technical manuals available from the publishers of each test, and talking to many of the publishers themselves, information contained in Buros' Sixth Mental Measurement Yearbook was reviewed. We also reviewed a report prepared by the UCLA Center for the Study of Evaluation entitled "Elementary School Evaluation Kit" which rated and ranked most of the available standardized achievement tests.

The Metropolitan Achievement Test, which was developed quite recently and normed on a large national sample, was used for evaluating the impact of the experiment. With the exception of grade one, three different payment tests were used in each grade. Three tests per grade were used for payment purposes in order to minimize problems with "teaching to the test." In selecting the payment tests, we attempted to find tests which were highly similar in terms of content and which were normed on comparable samples. Correlations between the evaluation test and each of the payment tests, based on the pre- and post- test administrations to the experimental groups, are included as Appendix C.

The decision to use norm referenced standardized achievement tests which measure what is generally taught was based on three considerations. First, we felt that these tests were quite acceptable in terms of the content covered, and that they provided a fair test of the contractors' programs (as indicated by the fact that most of the tests we used were among the tests recommended by the contractors). Secondly, in terms of the technical considerations (such as reliability), given the large number of tests we intended to use, the available nationally normed standardized achievement tests were superior to other types of tests. Finally, familiarity of school personnel with these tests helped reduce problems with test administration and interpretation of test results.

Reliability was a very important consideration in test selection, and every effort was made to select the most reliable tests from those available. The reliability data reported by the publishers for each of the tests used is included as Appendix B. In addition, we felt that careful selection of test levels and the use of composite skill area scores would provide maximum reliability for our students.

Composite Scores

The mathematics and reading tests in each of the batteries used are composed of several subtests. For example, the mathematics section of the seventh grade evaluation test is composed of three subtests: computation, concepts and problem solving. The publisher reports a KR_{20} reliability coefficient[5] of .89 for computation, .90 for concepts, and .91 for problem solving. If all three tests are used as a single test, however, the reliability (reported by the publishers and based on the same sample) increases to .96. The use of composite tests, then, provided a more reliable test than would the use of a single subtest.

Test Level Selection

Choosing appropriate levels of the tests was a complex matter. One of the major problems in assessing the impact of remedial programs on low achieving students is finding an accurate starting point or floor for the pre-test and, at the same time, allowing for enough growth during the year. Schools typically administer a test which is appropriate for the grade level in which a child is enrolled. For example, a student in the ninth grade is usually given a ninth grade test which might have a score of 5.5 as a floor. The student could score 5.5 no matter how little he knows and his "real" floor could be 3.0. In this case, the student's reading pretest score would be inflated by 2.5 years, which would result in lower payment to a performance contractor and would bias the evaluation itself. If the experiment included students with a cross section of abilities, selecting tests at grade level would have been appropriate. Administering different tests for each child at his own grade level was another possibility, but that approach was rejected for two reasons. First, giving different tests to all children at each grade level would have been administratively infeasible--25,000 children were tested on the evaluation instruments. Secondly, we felt there would be serious problems with the lack of comparability between scores or gains from different test levels, primarily because of content differences among the different levels.

---

[5] Based on grade six fall standardization group.

In order to reduce the dimension of test difficulty for low achieving students and to get as accurate a floor as possible without violating principles of test selection and administration, whenever possible, students in a given grade were matched with a level normed on the preceding grade. For example, in testing the seventh grade students, we attempted to select test levels which would be appropriate for sixth graders. This procedure, however, created a special problem in the selection of tests for first grade students, and to some extent with second graders because of the scarcity of standardized achievement tests which were normed on the performance of kindergarten and entering first grade students and which report grade equivalent scores. It was possible to select only two tests for first grade students: one achievement test for payment purposes and one readiness test which was used only for evaluation. Our reports from the various schools indicate that even the first grade achievement test was much too difficult for the study population in the pretest. Table I shows the test levels used in each grade as well as the publishers' recommendations concerning the grades in which each level should be used.

Because we used lower than grade level tests, there were a few cases of students "topping out," or receiving extremely high scores on the pre-test, leaving little room for improvement on the post-test. A far more serious problem occurred with respect to students "bottoming out" or receiving extremely low pre-test scores. Many students scored quite low on the pre-tests and to a lesser degree on post-tests. Because extremely low scores on tests of this type are generally considered to lack reliability, the contractors viewed these low scores as a sign of the instability and inappropriateness of the tests levels themselves. As a consequence of these concerns we computed the reliability coefficients for the evaluation test. Using the KR_{21} formula

TABLE I.

TEST LEVEL

Test/Level	Used with Grades	Recommended for Grades	Normed on Grades	G.E.'s Available for Grades
MAT - 1970				
Primary I	2	1.5 - 2.4	1.7, 2.1	1.0 - 6.0
Primary II	3	2.5 - 3.4	2.7, 3.1	1.0 - 7.0
Intermediate	7	5.0 - 6.9	5.1, 5.7, 6.1, 6.7	1.0 - 8.0
Advanced	8,9	7.0 - 9.5	7.1, 7.7, 8.1, 8.7, 9.1	2.0 - 9.9
California				
Level I	1,2	1.5 - 2	1, 2	.6 - 8.9 (8.7)
Level II	3	2 - 4	2, 3, 4	.6 - 13.6 (12.3)
Level IV	7,8,9	6 - 9	6, 7, 8, 9	.6 - 13.6
MAT 59				
Primary I	2	Last half 1	2.1	(1.0) - 3.9+
Primary II	3	2	3.1	(1.0) - 4.9+
Survey of Prim. Read. Development				
Forms A & B	2	1,2,first half 3	1, 2, 3	1.0 - 4.0
Forms C & D	3	2, 3, 4	2, 3, 4	1.0 - 5.0
SRA				
Level 1-2	2	end 1 - mid 2	1, 2, 3	(1-) - 4+
Level 2-4	3	end 2 - mid 4	2, 3, 4, 5	(1-) - 6+
Stanford				
Primary I	2	mid 1 - mid 2	1, 2	(1.0-) - 5.5+
Primary II	3	mid 2 - end 3	2, 3	(1.0-) -7.5+
CTBS				
Level 3	7,8,9	6, 7, 8	6, 7, 8	2.0 - 12.9
ITBS				
grade 7	7	7	7	2.2 - 12.3
grade 8	8, 9	8	8	2.4 - 12.9

for computing reliability the coefficients were quite high, on the order of .9 (see Appendix D for reliability coefficients). We believe that these high coefficients indicate that responses were not random, but that the results represent a reasonably accurate measurement of the students' abilities on the specific skills in the test.

Test Administration

To obtain test data permitting comparisons among the various groups of students being tested, it was essential that test conditions be carefully controlled and consistent from site to site. The test administration design specified:

1. The schedule to be followed in testing (including the sequence in which the tests were to be given).
2. How large the groups tested were to be.
3. Who was to administer and score the tests.
4. How the test examiners were to be trained.
5. What measures were to be taken to ensure test security.

Testing was supervised at each site by a representative of the test and analysis contractor (test coordinator). Examiners for the elementary grades were to be recruited from among certified substitute teachers. Guidance counsellors were to be used at the junior high school level. All examiners were to be given instruction in test administration by the test coordinator.

The testing schedule for the various grades was designed to minimize the effects of fatigue and limited attention span. First graders were to be tested in groups of 25 or less; second and third graders in groups of 35 or less; and seventh, eighth and ninth graders in groups of 100 or less with one proctor for every 50 students.

Payment tests were given only to the experimental groups. With the exception of first grade, three payment tests were administered in each grade, with one third of the students taking each test. During post-testing, each student took an alternate form of the same payment test he took during pretesting. The use of three tests in each grade was considered essential in minimizing problems with "teaching to the test."

In addition, in both pre- and post-testing, evaluation tests were administered before the payment tests. This was done to avoid introducing "practice effect" as a source of bias in the overall evaluation.

As is to be expected in a large-scale testing program of this type, there were logistical problems and some difficulty with student discipline. During pretesting incidents which could potentially affect validity of test results occurred in some grade groups at 10 of the 20 sites. At two of these sites, reservations about the validity of the results led to the decision to retest some groups of students, and conditions were greatly improved during the second testing. At the remaining sites, the incidents reported were judged not to be so extreme as to require retesting. The possible problems with validity for these grade/groups were explicitly noted and considered in the test and evaluation contractor's analysis. Some problems also occurred during post-testing, but these were generally less severe than pretesting problems. Retesting was not deemed essential in any of the cases. The data for all grade-groups was judged to be sufficiently valid to include in the analysis. As with the pretest, post-testing incidents which might affect validity are noted in the contractor's final report.

APPENDIX A

RELATION OF RAW SCORES TO GRADE EQUIVALENTS FOR TEST USED IN THE OEO EXPERIMENT

Grade 1 -- Reading (CAT '70)

An increase from to in grade equivalents represents an increase from to in raw score points.

Grade equivalents	Raw score points
.0 - .6	0 to 58
.6 - 1.0	58 to 63
1.0 - 2.0	63 to 80
2.0 - 3.0	81 to 104
3.0 - 4.0	105 to 112

Grade 2 -- Math (CAT '70)

Grade equivalents	Raw score points
.0 - .6	0 to 31
.6 - 1.0	31 to 37
1.0 - 2.0	38 to 60
2.0 - 3.0	62 to 80
3.0 - 4.0	80 to 87

Grade 2

An increase from to in grade equivalents represents
an increase from to in raw score points.

Reading

Grade Equivalents	Raw Score Points			
	Test I (ETS)	Test II (CAT)	Test III (MAT '58)	Test IV (MAT '70)
1.0 - 2.0	29-49	63-80	10-77	10-51
2.0 - 3.0	50-69	81-104	81-109	56-74
3.0 - 4.0	70-88	105-112	109-116	74-77

Math

Grade Equivalents	Raw Score Points			
	Test I (SAT)	Test II (CAT)	Test III (SRA)	Test IV (MAT '70)
1.0 - 2.0	9-41	38-60	22-51	13-43
2.0 - 3.0	42-58	62-80	53-74	44-55
3.0 - 4.0	58-62	80-87	75-88	55-60

Grade 3

An increase from to in grade equivalents represents an increase from to in raw score points.

Reading

Grade Equivalents	Raw Score Points Test I (CAT)	Test II (MAT '70)	Test III (ETS)	Test IV (MAT '70)
1.0 - 2.0	26-35	5-42	16-44	9-30
2.0 - 3.0	36-59	46-85	46-72	31-68
3.0 - 4.0	61-73	88-107	73-90	69-79
4.0 - 5.0	73-78	----	90-100	79-81
5.0 - 6.0	78-81	----	----	81-82

Math

Grade Equivalents	Raw Score Points Test I (CAT)	Test II (SRA)	Test III (SAT)	Test IV (MAT '70)
1.0 - 2.0	35-53	8-20	3-21	16-41
2.0 - 3.0	54-77	21-40	23-44	42-73
3.0 - 4.0	79-103	41-53	46-66	76-91
4.0 - 5.0	103-111	53-70	67-86	91-99
5.0 - 6.0	111-116	70-81	87-99	99-103

Grade 7

An increase from to in grade equivalents represents

an increase from to in raw score points.

Reading

Grade Equivalents	Raw Score Points Test I (ITBS)	Test II (CTBS)	Test III (CAT)	Test IV (MAT '70)
3.0 - 4.0	10-23	19-25	20-25	23-34
4.0 - 5.0	23-21	25-32	25-30	34-46
5.0 - 6.0	31-41	32-41	30-37	47-58
6.0 - 7.0	41-56	41-49	37-44	59-70
8.0 - 9.0	76-94	57-64	52-58	77-81
9.0 - 10.0	94-107	64-69	58-64	81-84
10.0 - 11.0	107-117	69-74	64-69	84-86

Math

Grade Equivalents	Raw Score Points Test I (ITBS)	Test II (CTBS)	Test III (CAT)	Test IV (MAT '70)
3.0 - 4.0	2-11	20-28	15-21	19-31
4.0 - 5.0	11-17	28-36	21-27	32-49
5.0 - 6.0	17-23	36-47	27-35	49-66
6.0 - 7.0	23-31	47-58	35-43	67-81
7.0 - 8.0	31-39	58-69	43-54	81-89
8.0 - 9.0	39-50	69-79	54-64	90-97
9.0 -10.0	50-60	80-88	64-72	97-101
10.0 -11.0	60-69	88-92	72-77	101-103

Grades 8 and 9

An increase from to in grade equivalents represents an increase from to in raw score points.

Reading

Grade Equivalents	Raw Score Points Test I (ITBS)	Test II (CTBS)	Test III (CAT)	Test IV (MAT '70)
4.0 - 5.0	19-29	25-32	25-30	26-34
5.0 - 6.0	29-37	32-41	30-37	34-41
6.0 - 7.0	37-46	41-49	37-44	41-50
7.0 - 8.0	46-60	49-57	44-52	50-58
8.0 - 9.0	60-75	57-64	52-58	59-65
9.0 -10.0	75-92	64-69	58-64	65-71
10.0 -11.0	92-105	69-74	64-69	71-74
11.0-12.0	105-117	74-77	69-73	74-78
12.0 - 13.0	----	77-80	73-76	78-81

Math

Grade Equivalents	Raw Score Points Test I (ITBS)	Test II (CTBS)	Test III (CAT)	Test IV (MAT '70)
4.0 - 5.0	7-14	28-36	21-27	25-36
5.0 - 6.0	14-19	36-47	27-35	36-46
6.0 - 7.0	19-25	47-58	35-43	47-60
7.0 - 8.0	25-31	58-69	43-54	60-71
8.0 - 9.0	31-40	69-79	54-64	72-82
9.0 -10.0	40-48	80-88	64-72	83-89
10.0 -11.0	48-58	88-92	72-77	89-93
11.0 -12.0	58-68	92-94	77-80	93-97
12.0 -13.0	-----	94-96	80-82	97-101

All data based on form of test used for pre-testing on the experiment.

APPENDIX B

RELIABILITY AND STANDARD ERROR OF MEASUREMENT OF TESTS USED IN THE OEO EXPERIMENT

SRA ACHIEVEMENT SERIES

	KR_{20}	S.E.M.
Level 1 - 2, Form C		
Total Arithmetic	.96	5.06
Level 2 - 4, Form C		
Total Arithmetic	.92	3.81

KR_{20} estimates are based on a stratified sample (N=200) drawn from the norm sample. Level 1-2 sample composed of beginning grade 2 students; Level 2-4, beginning grade 3 students. S.E.M. reported in raw score points.

STANFORD ACHIEVEMENT TEST

	KR_{20}	S.E.M.
Primary I		
Arithmetic	.95	3.18
Primary II		
Arithmetic Concepts	.91	2.88
Arithmetic Computation	.88	2.09

Reliability coefficients based on a random sample (N=1000) drawn from the standardization sample. Primary I sample composed of grade 1 students; Primary II, grade 2 students. S.E.M. reported in raw score points.

MAT '58

	r_1I	S.E.M.
Primary I		
Word Knowledge	.90	2.3
Word Discrimination	.87	2.5
Reading	.92	2.7
Primary II		
Word Knowledge	.93	2.2
Word Discrimination	.88	2.3
Reading	.94	2.8

Reliability coefficients are medians of four independent estimates of correlated split-half coefficients. Each estimate is based on a random sample (N=100) from a single school system. The four school systems were used to typify high, low and average performance. Primary I sample is composed of grade 2.1 pupils; Primary II sample grade 3.1. S.E.M. is reported in raw score points.

ETS SURVEY

	r_1I	S.E.M.
Forms A and B		
Reading	.909	
Forms C and D		
Reading Form C	.85	6.92
Reading Form D	.87	6.96

Form A coefficient computed by split-half method (N=304). Form C and D coefficients KR-#21 estimates based on grade 2 samples. S.E.M. reported in raw score points.

CALIFORNIA ACHIEVEMENT TEST

	KR_{20}	S.E.M.
Level I		
Reading	.950	4.18
Mathematics	.956	3.86
Level II		
Reading	.959	3.76
Mathematics	.953	4.28
Level IV		
Reading	.934	3.93
Mathematics	.930	4.10

Data derived from a sample (N=350 to 400) drawn from the standardization population and including students from each of the seven regions of the United States. S.E.M. reported in scale score units. Level I sample composed of grade 1.6 students; Level II, grade 2.6, Level IV, grade 6.6

IOWA TEST OF BASIC SKILLS

	r_1I	S.E.M.
Grade 7		
Vocabulary	.91	3.0
Reading	.93	3.7
Arithmetic Total	.90	2.1
Grade 8		
Vocabulary	.90	3.0
Reading	.93	4.0
Arithmetic Total	.91	2.1

Split-half reliability estimates based on a sample (approximately 12.5%) drawn from the standardization sample. Grade 7 sample composed of 2,723 grade 7 students; grade 8, 2,803 grade 8 students. S.E.M. reported in raw score points.

COMPREHENSIVE TEST OF BASIC SKILLS

	KR_{20}	S.E.M.
Level 3, Form Q		
Total Reading	.94	4.03
Total Math	.95	4.33

Estimates based on a sample (N=425) of grade 6.6 students drawn from the total standardization sample. S.E.M. reported in raw score points.

MAT '70

	KR_{20}	S.E.M.
Primary I		
Total Reading	.96	2.8
Total Math	.96	2.4
Primary II		
Total Reading	.96	3.1
Total Math	.95	3.5
Intermediate		
Total Reading	.96	3.6
Total Math	.96	4.0
Advanced		
Total Reading	.95	3.8
Total Math	.95	4.2

All reliability data is for Form G, based on all pupils in the fall standardization program. Primary I sample composed of grade 2.1 students, Primary II, grade 3.1; Intermediate, grade 6.1, and Advanced, grade 7.1 S.E.M. reported in raw score points and based on corrected split-half reliability estimates.

APPENDIX C

TEST CORRELATIONS

Table Correlations Between Evaluation Test and Certification Tests

Grade 2

Tests	Sample Size	Correlation Pre	Correlation Post
Reading			
MAT-CAT	332	.639	.825
MAT-MAT'58	370	.823	.904
MAT-ETS Survey	251	.651	.787
Math			
MAT-CAT	367	.751	.755
MAT-SRA	348	.779	.792
MAT-SAT	314	.729	.835

Grade 3

Tests	Sample Size	Correlation Pre	Correlation Post
Reading			
MAT-CAT	433	.764	.867
MAT-MAT'58	358	.799	.900
MAT-ETS Survey	148	.585	.697
Math			
MAT-CAT	419	.749	.828
MAT-SRA	344	.729	.835
MAT-SAT	320	.642	.834

Grade 7

Tests	Sample Size	Correlation Pre	Correlation Post
Reading			
MAT-CAT	342	.770	.779
MAT-CTBS	362	.726	.790
MAT-ITBS	304	.601	.786
Math			
MAT-CAT	305	.707	.799
MAT-CTBS	352	.773	.832
MAT-ITBS	276	.467	.651

Grade 8

Tests	Sample Size	Correlation Pre	Correlation Post
Reading			
MAT-CAT	347	.794	.807
MAT-CTBS	319	.793	.826
MAT-ITBS	299	.797	.745
Math			
MAT-CAT	315	.801	.839
MAT-CTBS	286	.748	.841
MAT-ITBS	278	.501	.687

Grade 9

Tests	Sample Size	Correlation Pre	Correlation Post
Reading			
MAT-CAT	318	.830	.863
MAT-CTBS	330	.816	.854
MAT-ITBS	291	.762	.824
Math			
MAT-CAT	315	.890	.878
MAT-CTBS	336	.664	.878
MAT-ITBS	281	.528	.710

APPENDIX D

TEST RELIABILITIES

ESTIMATED RELIABILITY COEFFICIENTS (KUDER-RICHARDSON 21) OF THE PRE- AND POSTTEST FORMS OF THE EVALUATION READING AND MATHEMATICS FOR EACH GRADE[a]

	Reading		Mathematics	
	KR-21	N[b]	KR-21	N[b]
Grade 1				
Pretest (SEAT, Level I)[c]	0.94	2139	0.94	2124
Posttest (CAT, Level I, Form B)	0.90	2139	0.92	2124
Grade 2				
Pretest (MAT, Primary I, Form F)	0.92	2702	0.88	2531
Posttest (MAT, Primary I, Form G)	0.98	2702	0.97	2531
Grade 3				
Pretest (MAT, Primary II, Form F)	0.94	2482	0.98	2357
Posttest (MAT, Primary II, Form G)	0.96	2482	0.94	2357
Grade 7				
Pretest (MAT, Intermediate, Form F)	0.93	2319	0.93	2286
Posttest (MAT, Intermediate, Form G)	0.94	2319	0.95	2286
Grade 8				
Pretest (MAT, Advanced, Form F)	0.91	2256	0.90	2153
Posttest (MAT, Advanced, Form G)	0.92	2256	0.93	2153
Grade 9				
Pretest (MAT, Advanced, Form F)	0.93	2089	0.94	2077
Posttest (MAT, Advanced, Form G)	0.93	2089	0.94	2077

---

[a] The sample used to estimate KR-21 were full-year students with both a pre- and posttest score in the appropriate subject.

[b] N = the number of students in each sample.

[c] See an earlier section of this report for more complete identification and discussion of each test.

Chapter III

CONTRACTUAL PROCEDURES

by

Charles Stalford

INTRODUCTION

This paper discusses the role in the remedial education experiment of performance contracts--the theory underlying them, their structure and difficulties in administering them, and the final settlement process.

The performance contracts were a major aspect of the educational programs tested. While all or parts of the experimental programs had been used previously, none had been scientifically tested on a performance basis.[1] The performance contract was envisioned as a major step to increase the effectiveness of the experimental programs. It was hoped it could do so through one or both of two ways:

-- Because payment in a performance contract is based upon results, clear measurement of educational goals could be expected and all parties held accountable for results.

-- The contracts contained incentive provisions which required achievement of minimum results before payment was made and rewarded achievement beyond the minimum. Therefore, contractors were encouraged to perform at maximum effectiveness, a level which it was hoped would exceed regular school programs.

In order for a performance contract to be effective, its provisions must be unambiguous and the terms of payment clearly described. Similarly, the contract should be genuine; that is, one in which payment is based only upon the specified outcomes and is not prevented or provided on the basis of loopholes, hidden provisions, or faulty measurement specifications. Under an ideal contract, motivation to perform is maximized.

---

[1] A performance contract had been used in the 69-70 school year in Texarkana, Texas. This project, funded by the U. S. Office of Education, was essentially a demonstration; there was no control group and no structured evaluation design.

As is frequently true with early attempts to implement new techniques, the performance contracting experience in this project was less than ideal. In the discussion that follows, the reader is invited to keep one question in mind: Were contract procedures used in such a way as to retain the incentive aspect of the basic experimental hypothesis? The answer to that question is a major factor in the overall evaluation of the experimental outcomes.

This paper is organized into four sections: The first discusses general contractual relationships between the OEO, its support contractors, the school districts, and the private firms. The second discusses the incentive structure of the performance contracts and the methods used to derive actual payments to the private firms. The third section discusses specific provisions of the contracts between OEO and the 18 school districts and between the districts and the six private firms; problems that arose during the school year in implementing these provisions; and the manner in which problems were handled during the two phases of subcontract renegotiations. Finally, we draw some conclusions and implications that might be considered in drafting future performance contracts.

CONTRACTUAL RELATIONSHIPS IN THE EXPERIMENT

Before discussing the contracts themselves, it is essential to understand the relationship of the parties in the experiment. The 18 school districts signed contracts with OEO in which each agreed to participate in the experiment with a designated education technology company. The contracts between the private companies and the school districts, technically then, were subcontracts, although the firms frequently are referred to as contractors. In addition, OEO had direct contractual relationships with Education Turnkey Systems, Inc., of Washington, D. C., the management support contractor, and the Battelle Memorial Institute of Columbus, Ohio, the testing and evaluation contractor. These relationships are illustrated in Figure I.

The Support Contractors

Specific tasks assigned to Education Turnkey included:

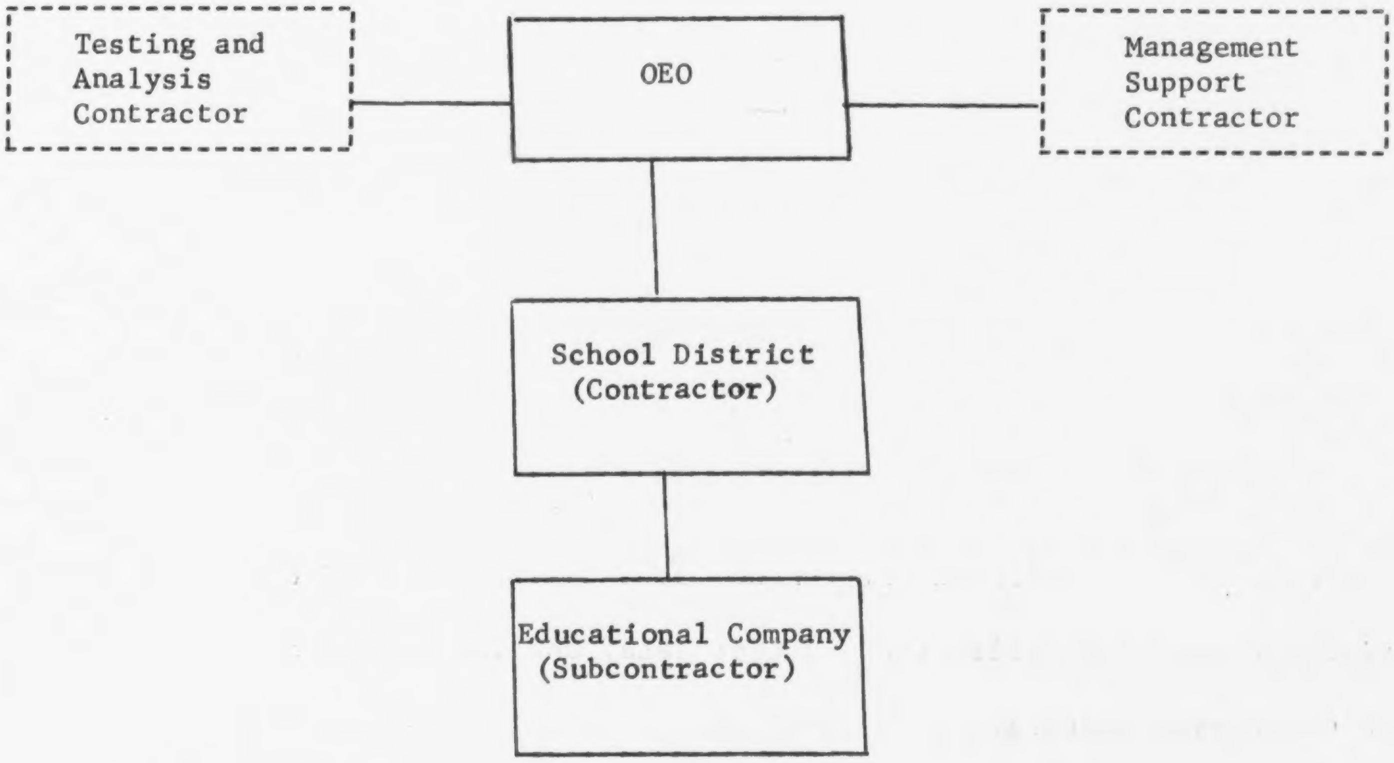
-- Assisting the OEO staff in selecting the participating school districts and identifying and selecting the specific schools and student populations that would be included in the experimental and control groups.

-- Establishing a system to monitor and document the operations of all experiment sites and private firms.

-- Assisting each school to ensure that contract provisions were being met.

Figure I

Contracting Relationships in the Experiment


Testing and Analysis Contractor
OEO
Management Support Contractor
School District (Contractor)
Educational Company (Subcontractor)

- -- Examining all curricular materials used by the private firms to ensure that teaching to the test did not occur.
- -- Developing measures of cost/effectiveness and collecting data on school district and subcontractor costs.
- -- Assisting OEO in identifying and discussing policy issues relating to the experiment.

The Battelle Memorial Institute was responsible for advising OEO in the selection of standardized tests administered to the control and experimental students for evaluation purposes and to the experimental group for use as the basis for determining the private firms' payments. In addition, Battelle was responsible for administering all pre- and post- tests, for certifying the relevancy of the criterion referenced tests administered by the private firms, for collecting the test data, and for analyzing the experiment results.

The Prime Contractors

As noted above, the 18 school districts served as prime contractors in this experiment. In the statement of work included in each prime contract, the schools agreed to:

- -- Enter into a subcontract with their assigned subcontractors.
- -- Cooperate with the management support contractor and the testing and analysis (evaluation) contractor.
- -- Not enter into other performance contracts involving the experimental or control group students.

- -- Hire a full-time professional project director and an assistant to serve as representatives and liaison with all parties in the experiment.
- -- Provide office and classroom space to the subcontractor.
- -- Enroll the children designated by OEO in the experiment and control classrooms.
- -- Facilitate test administration.
- -- Provide all data needed by Education Turnkey and Battelle.
- -- Provide general support in dealing with the community, parents, and teachers.
- -- Examine operating procedures and modify them if they would conflict with the experiment.

The last provision asserted the primacy of the experimental requirements over normal school procedures. For example, most districts had already established testing schedules for their students, but the experiment prohibited the administration of any standardized tests outside the experiment to either the experimental and control group students. Thus, except in those cases where state law required that the school tests be given, the schools dropped their own testing requirements.

Funds were allotted to the schools as reimbursement for their own administrative expenses occasioned by the experiment and to pay the subcontractors. The OEO did not pay the private firms directly. School expenses were reimbursed on a cost basis. Table I lists the contract amounts for each district, subdivided into allocations for school administrative expenses and subcontract payment.

School expenses were principally the salaries of the project director and his assistant; a secretary's salary, fringe benefits; overhead and necessary travel. In addition $3,200 was provided for each site to allow the firm to refurbish its classrooms. With one exception, all other direct costs of personnel, supplies, materials, and so on were borne by the private firms. The exception was Alpha Learning System, in whose programs certified teachers remained employees of the school district and their salaries were paid from the schools' administrative allotment. In all other programs, instructional personnel were employees of the companies and were paid from company funds. (Alpha hired additional paraprofessionals directly.)

OEO-Subcontractor Relationships

While no direct contractual relationships existed between OEO and the private firms, OEO retained rights of approval and exercised substantial control over the subcontracts. Several other noncontractual aspects of the experiment also combined to create a strong direct relationship between OEO and the subcontractors. First, the education firms were selected by OEO before the school

Table I

Total Prime Contract, Subcontract Ceiling Price, And School Administrative Allotments By School District

School District	Total	Subcontractor (Ceiling)	Administrative
Anchorage	$355,282	$272,200	$ 83,082
Athens	303,020	242,100	60,920
Bronx	343,046	288,000	55,046
Dallas	300,667	252,000	48,667
Fresno	300,265	240,000	60,265
Grand Rapids	323,714	180,000	143,714[a]
Hammond	343,778	288,000	55,778
Hartford	321,823	180.000	141,823[a]
Jacksonville	343,550	288,000	55,550
Las Vegas	299,994	240,000	59,994
McComb	264,335	223,200	41,135
Philadelphia	297,541	240,000	57,541
Portland	309,434	264,000	45,434
Rockland	300,461	252,000	48,461
Seattle	345,050	283,800	61,250
Selmer	287,541	242,100	45,441
Taft	245,001	153,000	92,001[a]
Wichita	295,290	242,100	53,190
TOTALS	$5,579,792	$4,370,500	$1,209,292

---

[a] Alpha Learning System programs, in which direct costs of certified teachers' salaries were included in school district administrative allotments. Paraprofessionals were on Alpha's payroll.

districts were selected for the experiment. Second, the school districts and companies were matched by OEO to ensure the appropriate mix of site and program characteristics necessary for the experiment design. Third, OEO and its management support contractor drafted the model subcontract used as the basis for the actual subcontracts.

Further, contract provisions and changes were negotiated directly by OEO with the subcontractors; however, school district representatives were party to all original negotiation sessions. (Part of the original subcontract negotiations in fact took place in the local school districts.) Later, renegotiations were held directly between OEO and the subcontractors. The school districts were a signator to all final agreements with the subcontractors, however.

The Nature of Performance Contracts

In any formal contract, the terms and conditions of the agreement are reduced to writing, the parties are obliged by the signatures to carry out the agreement and can be held legally responsible for default, and the payment usually is given to one party by the other in consideration for carrying out the contract requirements.

Contracts frequently are characterized by their intended purpose and the payment criteria. The most preferred contracts are so-called firm fixed-price contracts, which usually are used when specifications for the item to be procured are fixed and the anticipated costs are well established. A contract to manufacture 1,000 standard typewriters at a specific price would be such an agreement. Once signed, the price of the fixed price contract is binding upon the contracting agency and its contractor. The latter benefits from the prospect of higher profit margins if he can reduce costs effectively. The former benefits from an established price.

The past 20 years have seen widespread Federal use of cost-reimbursable contracts. In these procurements, the specifications of the product may be less certain and the risk involved in meeting them greater. Here, a reimbursable agreement is signed in which a target cost is set and the contractor is reimbursed for allowable, allocable and reasonable costs up to that amount. Unlike the firm fixed price contract, the cost contract provides for the contractor's books to be audited to substantiate his claims for costs incurred.

The cost-plus fixed fee contract (CPFF) adds a specific fee, usually expressed as a percentage of anticipated costs. If higher-than-anticipated costs are incurred on a cost contract, the contractor may be paid for them if funds are available, and if the contracting agency approves of the reimbursement. Much Federal research and development, where risks are high and product specifications uncertain, has been on a cost basis.

Because contracts usually require performance of some kind, such as delivery of a product, most contracts could properly be called performance contracts. For example, the specifications for an item, such as number, size, and weight must be met before payment is made.

Incentive contracts, similar to those used in the experiment, have been a relatively recent development in Federal contracting. They have been used principally by the Department of Defense and NASA, although the Department of Labor also signed an incentive contract governing part of the operations of a Job Corps Center. In a fixed price incentive contract (FPI), the entire payment will depend upon criteria related to the completion of the task, such as schedule dates, performance, or costs. For example, an incentive clause could provide a bonus for delivering the product before a certain date or for meeting certain quality standards.

In a second type of incentive contract, the cost-plus incentive fee (CPIF), the contractor is assured reimbursement for his costs, plus a fee, with the amount of the fee determined by performance criteria. The Job Corps CPIF contract, for example, called for payment of costs plus a fee whose size was determined by criteria such as the number of corpsmen who gained a high school equivalency diploma or who were placed in some job or school activity.

CPIF contracts also may include cost sharing between the government and the contractor. If cost is an incentive factor, payment to the contractor will vary according to his actual costs incurred. The contractor may be entitled to retain a share of the savings if costs are below the target, or if costs exceed the target he may be liable for a portion of the excess. A typical sharing ratio would be 80/20, in which case the government would pocket 80 percent of the savings (the contractor retaining 20 percent) and absorb 80 percent of any excesses.[2/]

The performance contracts used in this experiment were more similar to the FPI than CPIF contracts. The performance criterion was educational output, measured by various tests. A scale of educational gains necessary to earn various prices was set, but only a ceiling was established beyond which no payment would be made. There was no fixed price of any kind. The contractors were not assured reimbursement for any portion of their costs; that is payment was based only on educational gains, regardless of the contractor's cost experience. The contracts were negotiated at a level which OEO believed could afford up to 20 percent profit

---

[2/] For a brief summary of incentive fees in contracts, see "The Performance Contracting Concept in Education", The Rand Corporation, R-699/1-HEW, May 1971, pp. 55-66.

if the contractor was maximally successful. This rate of profit was considered equitable in view of the risks undertaken by the contractors.[3]

There were no time incentives in these contracts. All students were to remain in the experimental programs the entire school year, with gains measured at the completion of that time. This was unlike the original Texarkana contract, in which prices for specified gains rose in inverse proportion to the time spent in instruction. Such a time incentive in this experiment would have allowed and encouraged the subcontractors to discharge students throughout the year and cycle new ones into the program. This would have confounded the analysis plan, which was to compare experimental and control students after a full year of instruction by the different techniques.

Before proceeding, it is worth noting why incentive contracts in this experiment are particularly unusual.

First, schools generally have contracted for auxiliary educational services, such as lunch programs, bussing systems, and maintenance, among others. Instruction, however, is normally carried out directly by school systems, and therefore traditionally has not been performed under a contract.

---

[3] While there were no incentive provisions relating to cost sharing, the firms were required to renegotiate the price if they made any program revisions that substantially reduced their costs. This operated essentially as a disincentive to reduce costs. Considering the magnitude of the educational gains required, however, OEO believed the possibility of this clause being invoked was slim. It never was.

There are administrative and legal reasons for this, the principle one being that under state law a local school system is delegated responsibility for instruction of children from the state. It usually cannot redelegate that responsibility to a third party, such as a contractor. It has generally been held, however, that programs in which the school maintains control over the contracted educational programs are legal. The first unusual aspect of these and other performance contracts in education, therefore is not the incentive clauses, but the procurement of educational services by contract.[4/]

Guarantees are not new. Everyone has heard the phrase "satisfaction guaranteed or your money back," but that phrase has seldom been attached to an education program in the public schools. Educating children is undoubtedly a more unpredictable undertaking than mass-production of hardware. Most social programs are. Teacher organizations frequently state, for example, that they cannot be held solely "accountable" for student learning, to the exclusion of other factors such as malnutrition, lack of parental interest in education and broken home life, etc. Education technology firms, on the other hand, said they were willing to accept this responsibility.

[4/] For a fuller discussion of the legal aspects of this issue see Reed Martin, "Performance Contracting: Making it Legal," Nations Schools, January, 1971.

An additional unusual aspect of these contracts was the apparent degree of risk undertaken. Subcontractors generally claimed that before receiving even a minimum amount, they would produce greater gains than normally obtained by a school. (In some respects an equally unlikely occurrence might have been a private company funding a moonshot on a money back basis.) This will be discussed in detail in the following section.

THE INCENTIVE STRUCTURE

The structure of the incentive clauses was identical in all 18 subcontracts:

(1) All payments in the contracts were based upon individual student test results. Each child in the experiment was to be tested and a payment calculated for him in accordance with his test results and the incentive scales. The aggregate of such payments made up the total reimbursement to the contractor.

(2) All contracts stipulated a minimum guaranteed level of achievement before payment was made. The lowest such guarantee was 0.5 grade level equivalents (GLE), in the elementary grades. The highest was 1.5 GLE's in the secondary grades. The median overall was approximately 1.0 GLE's. A price was set for each student whose achievement improved to the guarantee level.

(3) Contractors were asked to specify the maximum gain, on average, they thought students in their programs could achieve (as shown in Table II), and a price was set for that "maximum average." The payment for each tenth of a grade level improvement between the maximum average and minimum guarantee was

Table II

Summary of Anticipated Average Maximum Gains by Subcontractor

Contractor	Grade-Subject	Average Maximum Gain (Grade Equivalent) On Standardized Tests
Alpha Learning Systems	Grade 1-3	1.7
	Grade 7-9	1.7
Learning Foundations	Grade 1-3	1.9
	Grade 7-9	2.2
Plan Education Centers	Grade 1 Math	1.0
	Grade 1 Reading	1.5
	Grade 2,3 Math	1.5
	Grade 2,3 Reading	2.0
	Grade 7-9 Math	2.0
	Grade 7-9 Reading	3.0
Quality Educational Development	Grade 1-3	2.0
	Grade 7-9	2.0
Singer-Graflex	Grade 1-2	1.5
	Grade 3,7-9	2.2
Westinghouse Learning		1.7

determined by dividing the difference between the maximum and minimum prices by the difference in the maximum and minimum grade level equivalents.

(4) Up to 25 percent of the contract ceiling price was based upon results of five interim performance objective (IPO) tests in each subject. These were criterion-referenced tests developed by the firms and oriented to their own curriculum objectives. They were scored on a pass-fail basis, with passing set at 75 percent on each. The subcontractor was paid 2.5 percent of the calculated ceiling price for each child each time he passed an IPO. Each firm separately negotiated its own combination of minimum guarantee price, incentive price and interim test prices.

The request for proposals from private firms established a general target of $200 per student per subject as the maximum price of the entire contract. The $200 figure was chosen in part to keep the contract price at a level that superintendents could consider for a future operational performance contracting project.

It later became obvious, however, that prices should vary to account for particular characteristics of subcontractor programs and incentive scales. Nevertheless the $200 figure was retained as a general target in subcontractor negotiations.

The incentive scales for each contractor are shown in Table III. They should not, however, be viewed as indicative of subcontractor costs, since nothing prevented a firm from spending above the maximum price to improve its chances of success.[5]

An hypothetical example of a payment calculation is:

Student A, in the Alpha program, Grade 1 tested as follows:

	Reading	Math
	Standardized Testing (GLE's)	
Post	2.0	1.7
Pre	0.9	1.0
Gain	1.1	0.7
IPO Results		
#1	75% (Pass)	70% (Fail)
#2	65% (Fail)	75% (Pass)
#3	85% (Pass)	85% (Pass)
#4	80% (Pass)	80% (Pass)
#5	90% (Pass)	65% (Fail)

In reading, Alpha received $56.25 because the student achieved 018; or the minimum guaranteed gain, plus 3 x $6.25 = $18.75 for the 0.3 GLE gain above 0.8, for a total of $75.00. Student A passed four of five reading IPO tests; therefore Alpha earned $7.50 x 4 = $30.00 (80 percent of $37.50 total IPO payment).

In math, Student A did not meet the 0.8 minimum GLE guarantee;

---

[5] For a full discussion of costs, see another paper in this volume and the final report of Education Turnkey Systems, Chapter 7.

therefore Alpha earned no payment for him on this basis. The student did pass three of five math IPOs. Alpha therefore earned 3 x $7.50 = $22.50.

Total Alpha earnings for Student A are:

Reading GLE	$75.00
Reading IPO	30.00
Math GLE	0
Math IPO	22.50
	$127.50

As shown in Table III, the maximum incentive price for Alpha is $150 per subject. Alpha therefore earned $105.00 ÷ 150.00 or 70 percent of the maximum price in reading and $22.50 ÷ 150.00 or 15 percent of the maximum price in math. For the two subjects combined, Alpha earned $127.50 ÷ 300.00 or 42.5 percent of the maximum price.

The maximum incentive price was used to establish the total contract price per site. But subcontractors' earnings for an individual student were not limited. If Alpha Student B in the first grade gained 2.2 GLEs in reading, Alpha would recover $56.25 for the first 0.8 GLE plus 14 x $6.25 = $87.50 for the 1.4 GLEs above 0.8. In total, $143.75 ($56.25 + $87.50) would be paid for reading GLEs. If Student B also passed all reading IPOs, an additional $37.50 would be earned, or $181.25 total in reading. The maximum incentive ceiling would apply only if Alpha were to achieve improvement at such a rate that the average payment for all students in all grades in both reading and math exceeded $300 (2 times the $150 ceiling for each subject). If the average payment were $140 in one subject and $160 in the other, Alpha would receive $300 per student. If it were $310 in one subject and $0 in the other, Alpha would forfeit $10 per student.[6]

The various components of the incentive structure interacted in such a way that subcontractors could earn the same amount of money for several different types of performance. To illustrate:

Westinghouse Basic Scale

1.0 minimum GLE - $75.00

0.1 GLE above minimum - $10.70

Price per student at average maximum of 1.7 GLE - $150

($75 + $10.70 x 7)

Case A:	100 students gain 1.7 GLE's =		$ 15,000
	(Average 1.7 GLE)	Total	$ 15,000
Case B:	50 students gain 3.4 GLE's =		$ 15,590
	50 students gain 0 GLE's =		0
	(Average 1.7 GLE)	Total	$ 15,590

---

[6] Conceivably, a subcontractor could have concentrated on reading to the detriment of math, or vice versa. However, under the incentive scale, the level of achievement required to make that financially rewarding would be prohibitive. For example, if a subcontractor had hoped to achieve 1.3 grade levels in both subjects, he would have had to approximately double that output in the chosen subject to offset the loss from the other. (Also, the subcontracts stated that instruction would be carried out for approximately 180 class hours in each of reading and math. Subcontractors occasionally spent more time in one subject than the other, but the difference was not great.)

Case C:	50 students gain 2.4 GLE's =	$ 11,240
	50 students gain 1.0 GLE's =	3,750
	(Average 1.7 GLE) Total	$ 14,990
Case D:	50 students gain 2.6 GLE's =	$ 12,310
	50 students gain 0.8 GLE's =	0
	(Average 1.7 GLE) Total	$ 12,310

The same phenomenon occurs with each subcontractor to a varying degree.

The subcontractor thus suffers severely for students who do not achieve the minimum guarantee. For each student who does not meet the guarantee, even though showing some gain, another student has to show substantial progress. It could be said that this incentive scale maximizes the incentive to achieve a homogeneous level of gains. It might also be legitimately claimed, however, that the incentive structures are too hard on the firms, and in retrospect, that some provision ought to have been made for gains below the minimum.

Summary of Subcontractor Incentive Scales (Per Subject) Table III

Contractor	(1) Minimum Guaranteed Gain (Grade Equivalent on Standardized Tests	(2) Price for Minimum Gain	(3) Price Per 0.1 Above Minimum Gain	(4) Maximum Price for IPO's	(5) Maximum Incentive Price
Alpha Learning Systems	0.8 (Gr. 1-3)[1]	$56.25	$ 6.25[2]	$37.50	$150.00[3]
	1.0 (Gr. 7-9)	75.00	5.36[2]	37.50	150.00[3]
Learning Foundations	1.0 (Gr. 1-3)	101.00	8.77	60.00	240.00
	1.0 (Gr. 7-9)	81.00	8.25	60.00	240.00
Plan Education Centers	0.5 (Gr. 1 Math)	50.00	20.00	50.00	200.00
	0.5 (Gr. 1 Read)	46.25	9.25	46.25	185.00
	1.0 (Gr. 2,3 Math)	50.00	20.00	50.00	200.00
	1.0 (Gr. 2,3 Read)	46.25	9.25	46.25	185.00
	1.0 (Gr. 7,9 Math)	50.00	10.00	50.00	200.00
	1.0 (Gr. 7,9 Read)	55.00	5.50	55.00	220.00
Quality Educational Development	1.0 (Gr. 1-3)	72.50	8.50	52.50	210.00
	1.5 (Gr. 7-9)	82.50	15.00	52.50	210.00
Singer-Graflex	0.5 (Gr. 1-2)	82.50	8.25	55.00	220.00
	1.0 (Gr. 3,7-9)	82.50	7.17	55.00	220.00
Westinghouse	1.0	75.00	10.70	50.00	200.00

NOTES - Prices shown are representative of all school districts for each contractor. If its prices varied by district, the lowest price is shown. Guarantee schedules for each contractor did not vary by district except where noted.

[1] - 0.5 minimum guarantee in Taft, Texas.

[2] - The actual price per 0.1 above the minimum was varied at different points in the scale. Figure shown is the average.

[3] - In the Alpha program, direct teacher salaries (excluding paraprofessionals) were paid by schools. All other contractor price figures include all direct instructional costs.

GENERAL CONTRACT PROVISIONS

Contractual relationships and procedures in the performance contracting experiment differed from what is "typical" for several reasons. First, because these were the first contracts of their kind, many problems that arose during the year were unanticipated; other anticipated problems that were taken into account when the contracts and subcontracts were written, turned out to be less or more severe than expected. Secondly, OEO had little "clout" or leverage over its prime contractors, the school districts, because they had little to lose if they did not or could not fulfill their contractual obligations. And, third, OEO had much more direct contact with the subcontractors than is usual.

In general, all parties in the experiment were conscientious in meeting their contractual responsibilities. When problems--such as underenrollment, missed tests, or lost instructional time--did occur, OEO initially refused to consider changing contract or subcontract provisions and instead sought to bring conditions into accord with the original provisions. This was not always possible, however, since many situations were clearly beyond the control of either the school districts, OEO, or the private firms.

As discussed in detail below, OEO re-entered negotiations with the subcontractors in February of the experimental year (1971), proposing a series of subcontract amendments to form the basis for the final settlement. The first phase negotiations continued through the remainder of the experimental year; the amendments were signed by the various firms between June and November. Until the very end of the negotiations, all parties, including OEO, were denied access to both the evaluation and payment test results. By late fall, however, this was no longer practicable for the OEO negotiators. By then, however, OEO's bargaining position was fairly well fixed, and therefore not affected by knowledge of the test results.

As soon as the amendments were signed, the subcontractors did receive the test results for their sites. They were uniformly disappointed and considered the earnings calculations, determined by the new amendments, to be unsatisfactory. The firms then submitted a series of additional matters that they felt justified further subcontract negotiations. These second phase negotiations have been completed with three of the six companies, and the final payments they received are shown in Table IV. But negotiations still continue with the remainder.

During both the first and second phase negotiations, OEO dealt individually with the subcontractors, but attempted to propose amendments that would be equitable to all. Details of the original contract provisions, problems that arose their implementation, and the methods adopted to deal with them in the first and second phase negotiations follow.

Table IV

Payments to Subcontractors[a]

	Payment	Subcontract Ceiling Price
Learning Foundations		
Bronx Dst. #9	288,000[b]	288,000
Hammond	207,176	288,000
Jacksonville	171,675	288,000
Plan Education Centers		
Athens	185,897	242,000
Selmer	242,100	242,100
Wichita	141,849	242,100
Westinghouse Learning Corp.		
Fresno	101,948	240,000
Las Vegas	127,266	240,000
Philadelphia	147,478[c]	240,000

[a] Payments were calculated by Programming Methods, Inc.

[b] Compensation based on reasonable costs to subcontractor.

[c] Compensation based on reasonable costs to subcontractor ($110,542) in the secondary grades

Student Enrollment and Attendance

The original subcontracts specified that:

- -- The school district would ensure that 100 (75 in three districts) children would be enrolled in each subject in each grade.
- -- Children who dropped out, for whatever reason, would be replaced within five days.
- -- After 20 hours of instruction, the private firms could request that a child be dropped only if he had been absent 10 consecutive days or 15 days in a three-month period.
- -- Any child leaving the program after more than 30 hours of instruction would be post-tested and his replacement pretested and post-tested. Payment for gains by drop-outs and their replacements was established by a separate incentive formula (discussed below).

The problem of replacing drop-outs, either those who moved from the district or those who were excessively absent, was more severe than anticipated, as was the problem of post-testing drop-outs. While a pool of potential replacements had been created before the experimental school year began, this pool was partially depleted at the very start of the experiment to replace students who had moved away from the districts over the summer. As the year progressed, replacement became more and more difficult. Toward the end of the year, it became virtually impossible, since parents were reluctant to enter their children into a nearly-concluded experimental program.

Thus, several methods were adopted to deal with the problems of underenrollment, missed post-tests, and replacements:

-- For children who had dropped out and not been post-tested and for children who had been in the program less than 30 hours, the private firms were paid an amount equal to the average payment for full-time students in each subject, prorated for the time the drop-out did remain in the program.

-- To compensate for underenrollment, the contractors' payments were calculated on the daily gains of students in the program throughout the year, multiplied times the number of student days lost in excess of the five-day replacement periods.

-- Only evaluation tests were administered to students who replaced drop-outs, because of the administrative difficulties involved in giving them two sets of tests. Similarly, when drop-outs could be found for post-testing, they were given only the evaluation tests. In both instances, the evaluation test results (rather than payment test results) were used to determine contractor payments.

-- Payment for children who dropped out but who had been given the evaluation post-test was based on a separate incentive schedule, which contained uniform prices for each tenth of a GLE gain. These prices were determined by dividing the payment for the average maximum gain by the price for achieving that gain. For example, the price for the average maximum gain, 1.7 grade levels, for Westinghouse was $150. This $150 was divided by 17, resulting in a drop-out/replacement price of $8.82 per tenth of a GLE. Before receiving any payment for drop-outs, however, the firms had to achieve skill improvements that would equal the minimum guarantee level on a projected basis. Thus, for example, if the minimum guarantee was 1.0, a student in the program had to improve by 0.5 GLE or more for the contractor to receive any payment.

Student Selection

The original subcontracts specified that under-achieving children were to be enrolled in the experimental and control groups. Using data from tests administered by the schools in the 1969-70 school year, OEO and its management support contractor selected the school (or where necessary, two or three schools) in each district with the lowest overall achievement test scores as the experimental school(s), and the school(s) with the next lowest overall achievement scores as the control school(s). Within each school, the children with the greatest combined deficiencies in reading and math were selected for the experiment. Although every attempt was made to screen out children who were mentally retarded or otherwise unable to benefit from the experimental program, a very few of these children were enrolled in it. During the first 20 hours of instruction, the firms could request that these "unqualified" children be dropped. These requests were infrequent, however, and the matter did not become an issue during the subcontract renegotiations.

A greater problem did develop with "over-achievers." In most districts, fewer than 10 percent of the students in the experimental group were found to be performing at or above grade level, but in one or two of the smaller districts, the percentage was higher. The contractors argued that their programs were designed for under-achievers, and therefore could not be effective with children who were at or above grade level. While some educators have suggested that the firms should have been able to achieve even better results with brighter children, the subcontract language was not sufficiently clear on this point and OEO accepted the firms' argument.

To adjust for over-achievement, OEO paid the firms for each tenth of a GLE these students improved, regardless of the minimum guarantee. The base price used to compute payment for each tenth of a GLE gain was either that used for student who dropped out adjusted by a factor proportional to the degree to which the student was above grade level or the price normally paid for each tenth of a GLE above the minimum guarantee, whichever was higher.

Time for Instruction

The original subcontracts specified that "a full academic year, consisting of approximately 180 class hours in each of reading and math," would be available for instruction.[7] The firms' guarantee and incentive scales were based on the assurance that this amount of time would be available.

As the experiment went into operation, however, several factors combined to decrease both the anticipated number of days and the number of minutes per day available for instruction, and this issue become pivotal to the second phase renegotiations.

First, the firms lost more time than they had anticipated because of pre- and post-testing. The subcontracts specified that pre-testing was to be completed "within the first ten days" of the school year and that post-testing was to begin "no earlier than ten days" before the end of the school year. The "no earlier than ten days" provision was renegotiated in the first phase discussions to 15 days, in large part because the firms did not want post-testing to take place during the last week of school. OEO's General Counsel interpreted this subcontract language to mean that the subcontractors should have expected 15 days to be lost at the end of the school year. But the General Counsel ruled that confusion was legitimate about the "within 10 days" clause for pretesting. Therefore, OEO stated that the subcontractors were entitled to a 165-day base for determining payments, rather than the originally specified 180 days.

---

[7] In some instances, the precise wording was "class periods" rather than "class hours."

The firms also lost time because of fire drills, teacher strikes, assemblies, and picture-taking sessions. These were disregarded, however, in estimating lost time, because it was felt these were normal school occurrences that should have been anticipated by the firms.

Secondly, a "class hour" frequently turned out to be nearer to 50 minutes than 60; in one instance, only 40 minutes were available to the contractor. Thus, in computing payments, adjustment was made for actual minutes available daily for instruction. Finally, since the contractors argued that their instructional time was also hindered by absences, "actual average attendance" (expressed as student days) was calculated and used as an adjustment factor.[8]

---

[8] Both of these figures were calculated as site averages, rather than for each grade and subject.

The factor used to adjust actual grade gains of each student, then, was determined by multiplying:

$$\frac{165}{\text{Actual Average Attendance}} \times \frac{60}{\text{Actual Class Minutes}}$$

Table V shows the results of these calculations for each site. The adjusted grade gain, on which payments were based, was calculated by multiplying the adjustment factor times a student's actual grade gain.

Testing

As noted earlier testing in the experiment was carried out both for evaluation and payment purposes. The original subcontracts, however, included provisions only for the payment tests; they were subsequently amended to indicate that a separate set of standardized tests would be used for the evaluation. This matter did not become a serious issue in the renegotiations, however.

Problems did occur because of lack of adequate provision for children who did not drop out of the program but who missed tests for one reason or another.

The original subcontracts provided only for missed post-tests (payment was to be based upon the average payment for students who had been post-tested) since it was felt that the firms were partially at fault for students' failure to attend testing sessions. As the year progressed, however, it became clear that the firms were not discouraging poor students from being tested, but rather that entirely

Table V

Table of Factors Used to Obtain Adjusted Grade Gains for Subcontract Payment

Site	Factor
Selmer	1.55
Dallas	1.39
Las Vegas	1.40
Anchorage	1.45
Athens	1.58
Wichita	1.52
Taft	1.37
McComb	1.46
Seattle	1.40
Grand Rapids	1.30
Hartford	2.10
Jacksonville	1.36
Rockland	1.36
Hammond	1.84
Portland	1.48
Fresno	1.40
Philadelphia	1.39[a]
Bronx	2.00[b]

---

[a] Compensation based on reasonable costs to subcontractor substituted for secondary grades

[b] Compensation based on reasonable costs to subcontractor substituted for all grades

natural reasons, such as sickness, were resulting in children's missing IPO tests, not completing or entirely missing standardized testing sessions, or being tested late. Under the original subcontracts, the firms could not recover payments for such cases. Consequently, the subcontracts were amended to provide that evaluation test results would be substituted for payment test results, whenever possible. If both tests were missed, and a makeup test could not be administered within 30 days, the results of students properly tested were to be substituted. For example, if a student missed both pretests, the mean evaluation score for the experimental group in his district's grade and subject was used as the payment premeasure. Similarly, the average payment for students taking IPOs was substituted for those who missed them. These provisions were used to calculate payments only for students who attended at least 75 percent of the regular class sessions.

About two-thirds of the students did remain in the program for the full year, and evaluation test scores were available for most of the replacements for those who dropped out. Thus, "not tested" students amounted to only about 10 percent of the total.

Another testing problem concerned students who scored at the ceiling of the grade level equivalent table on the post-test. The private firms raised the legitimate question of whether the students might not have been recorded as gaining more if the test ceiling had been higher. To deal with this problem in the primary grades, the contracts were amended to shift the basis for payment from the payment test to the evaluation tests, which had higher ceilings. If the student scored at the ceiling even on the evaluation test, or if the evaluation test ceiling was lower than payment test (as it was in the secondary grades), payment was based upon the average gain of all students properly tested, or the individual "topped out" gain, whichever was greater. This problem did not occur very frequently, however.

The First Grade Problem

While several tests were considered appropriate for the other grades, only one achievement test was found for first grade, and its grade equivalent table went only as low as 0.6. On the pretest, this was too difficult for most of the first graders. Consequently, an arbitrary pretest score was assigned as a basis for calculating pre/post-test gains. OEO initially suggested that 0.3 be used as the base, but some contractors argued that even this was unfair. OEO finally agreed to use 0.2 as the base for those children whose recorded pretest level was 0.6.

Payment Bonds

In order to meet the subcontractors' cash flow requirements during the school year, the subcontracts provided for seven provisional payments, totalling 80 percent of the maximum subcontract price, at intervals during the year. While the amounts paid were independent of amounts later earned under the incentive scales, they were tied to specific milestones, principally the administration of the IPOs. The payments represented an advance to which additional earned funds would be added if the final amount earned under the incentive clauses was more than the 80 percent; conversely, if the amount finally earned was less than the 80 percent, the firms were to return the difference to the government.

To protect the government against the risk of losing the advanced funds in the event that the firms did not ultimately earn 80 percent of the maximum subcontract price, the subcontracts required the firms to post a bond or provide other indemnification satisfactory to the government to insure against loss of funds. While the subcontracts initially specified "performance" bonds, payment bonds actually were required. (A performance bond usually is used to guarantee completion of a task as specified in a contract; a payment bond guarantees repayment to the contracting agency if the contractor defaults on its obligations.)

Implementing this provision was difficult for the firms and the OEO (and is likely to present problems to school districts undertaking performance contracts in the future). Because the firms involved in the experiment--and in performance contracting generally--are, for the most part, new and small, they found the payment bonds almost impossible to obtain. Only Westinghouse Learning Corporation, one of the two largest subcontractors, was actually able to obtain a payment bond. Three other firms pledged corporate stock or funds payable to OEO under their subcontracts. But two of the smaller firms were unable to make any satisfactory bonding or indemnification arrangements, despite repeated efforts, and proceeded without them.

Supplemental Instruction

The contracts specified that the school districts would not teach reading or math outside the experiment to students in the experimental classrooms since it was essential to the evaluation that their only reading and math instruction by in the performance contracting classrooms. This was not a problem in the secondary grades, but in the primary grades, reading and reading-related activities represent a substantial portion of the school day. OEO finally adopted a ruling that direct instruction in reading skills, vocabulary, word attack, and so on, was to be conducted only in the experimental classrooms; other normal supplemental activities, such as silent reading time and story telling, were not prohibited.

The Bronx and Philadelphia

In two districts, the Bronx and Philadelphia, the experiment was severely hampered by such severe obstacles that OEO was forced to settle not on a performance basis but rather partially or completely on an estimate of reasonable costs.

Controversy marred the Bronx experiment from the time it was announced. It was the first project of this type undertaken by the Community School Board, which had recently been established under the New York City Schools decentralization plan. The local teachers' union attacked the proposed program publicly just as classes were to begin, and continued its campaign through the news media throughout the year. The union contested the use of paraprofessionals in classrooms, the lack of union involvement in the contract negotiations, alledged disruption of a program for Spanish-speaking children, and many other factors. The Community Board answered these allegations, but extreme mistrust between union teachers and those involved in the experiment continued to hamper the program throughout the year.

In addition, disruptions and disorder during the pretesting sessions became so intolerable that the tests had to be suspended while an intense, three-week campaign for community support was undertaken by the school board. Instruction did not actually begin until October. Confusion in identifying and enrolling students in the program produced uncertain rosters of participants; maintaining accurate enrollment and attendance records also was a significant problem.

Absenteeism during both the pre- and post-testing sessions was high; many students who were present skipped all the questions or attended only part of the testing sessions. Consequently, less than half of the enrolled students in some grades had both a complete pretest and post-test in the same subject. In one instance, out of the intended 100 students, only 22 were enrolled.

In light of all these factors, OEO decided that attempts to fix the "blame" would be fruitless and inappropriate, and agreed to reimburse the contractor for reasonable costs, not to exceed the ceiling price of the subcontract. Costs exceeded the ceiling, so the firm was paid the full ceiling amount.

Similar problems plagued Philadelphia. Delays attributable to both the firm and the school district were encountered in enrolling students, and when school opened, the two disagreed as to which was responsible for providing various supplies and for completing refurbishments in the experimental classrooms. The firm was not completely satisfied with the equipping of its classrooms until November, and encountered early difficulties in gaining access to school buildings after hours of planning and logistics. All district schools opened late, and then were further disrupted in October by a brief teacher strike. By this time, the subcontractor was having difficulty maintaining discipline and providing instruction in the secondary grades. In addition, the firm's property was vandalized and stolen.

Because the problems were less severe in the primary grades than in the secondary grades, payment for those children was based on the original incentive formula. But a compensation of $110,542, reflected the subcontractor's reasonable costs, was agreed upon for the secondary grades.

CONCLUSIONS

The evaluation of this experiment was designed to show whether innovative programs carried out on a performance contract basis were more effective than regular school programs. In order for the test to be a true one, the contracting process had to offer real incentives to maximize achievement. As stated in the introduction, a critical view of the contracting experience in this experiment is necessary to judge the overall worth of the findings. The experience gained is also relevant to future performance contracts.

While the performance contract concept is simple, "You pay for what you get" etc., its execution for educational programs was shown by this experiment to be complex. The original subcontract signed by the companies was a comprehensive instrument which took note of various fiscal, legal, testing, and administrative factors. The incentive clauses appeared straightforward and rigorous. Yet a large number of amendments still were required to the contracts. Interestingly, no changes were made to the incentive clauses themselves. While the adjustment for lost instructional time had the effect of reducing the guarantee schedules, the concept of a minimum gain required for payment with incentive payments beyond that point was retained. Most of the amendments dealt with conditions surrounding the implementation of the programs.

The final settlement of these contracts by OEO has been a complex matter, in which the legal interpretation of clauses has guided settlement. As noted, ambiguous contract language sometimes made settlement more difficult. Where there has been divergence between the two, the contracts have been settled by the language of the contract rather than the assumed intent of the parties.

While continuous renegotiations extended over the last half of the school year and beyond, OEO made it clear that there was no intent to set aside the contracts as a basis for settlement.

In an ideal research experiment, a performance contract would have clearly defined incentive scales, so that behavior could be analyzed in terms of responses to them. For example, analyses might be conducted to determine whether a contractor sought to maximize gains with a few students or achieve minimum gains with all, or emphasized one set of grades at the expense of another. With the prolonged renegotiations and adjustments to the original subcontracts in this experiment, the ultimate terms of settlement were in some doubt during the school year. This would make research into the effects of specific incentive provisions somewhat unreliable. Nevertheless, with the exception of the two sites where cost-based adjustments were made, the structure of the incentive contract was maintained as the basis for settlement.

No agreements which substantially altered the original contracts were made until after the experiment was completed; therefore the basic thrust of a performance contract, which is to optimize performance in order to maximize reward, has been retained.

Much of the difficulty in administering these contracts has been related to the size of the experiment itself. The necessity for OEO to consider positions applicable to all contractors while dealing with each individually made contract administration cumbersome. A school district administering a single contract without the tripartite OEO-school-company relationship could expect to have an easier task. It would not have any lesser need for a clear contract, however.

In order for a performance contract to be a useful management tool, it should not be so difficult to administer that the value of educational benefits realized is submerged under contractual difficulties. Experience in this experiment has demonstrated the need to comprehensively define conditions under which the performance contract project is to be carried out and provide remedies for their breach. There is no reason why any future contract could not avoid many of the difficulties described here and be a more manageable tool.

On the other hand it is quite likely that difficulties in resolving these contracts would have been eased if the results had been more successful. In addition to clarifying the contractual document, another means to improve its usefulness might be to incorporate more modest performance provisions which did not require the magnitude of gains incorporated not only in these contracts but in others to date as well. It does not seem necessary for a contractor to achieve an average of two full grade levels per student in order to achieve maximum payment or impress the community. Given the present discontent with compensatory education programs, a performance contract project which achieved a full year's growth with most students at moderate cost would seem a reasonable goal to pursue.

159

Chapter IV

ANALYSIS OF PROGRAM COSTS

by

Charles B. Stalford

This paper is based upon work performed by the management support contractor in the experiment, Education Turnkey Systems, Inc., of Washington, D. C. The final report of that contractor contains a chapter which treats the subjects discussed here at greater length. Summaries of data in this paper are abstracted from that report, and Education Turnkey staff have provided additional assistance in preparing this paper. Interpretive conclusions are the author's.

INTRODUCTION

While the evaluation emphasis in the performance contracting experiment was on experimental/control achievement differentials, attempts also were made to examine cost differentials. The primary analysis was carried out by Education Turnkey Systems, the management support contractor, on the basis of data from the participating school districts and private firms. That analysis is described briefly here and in more detail in Education Turnkey's final report.[1]

Education Turnkey limited its analysis to a review of costs per year primarily because achievement results were not available for a more sophisticated cost-effectiveness analysis when its report was prepared. Since then evaluation findings have shown little difference between experimental and control achievement; therefore possible studies of cost-effectiveness would be limited to those relatively few instances where experimental programs demonstrated positive effects.[2]

The process of analyzing costs of educational programs in public schools is not very well developed, in part because of the character of most public school budgets. Most states' school codes require budgets to be organized along administrative functions

---

[1] For greater details on the methodology used, see Final Report to the OEO: Performance Incentive Remedial Education Experiment. (PB 202830), which is available for $3.00 from the National Technical Information Service, Springfield, Virginia 22151.

[2] Achievement results data are discussed in the Garfinkel/Gramlich paper, which is Chapter I of this volume and in Final Report on the Office of Economic Opportunity Experiment in Education Performance Contracting, (PB208947) Battelle Memorial Institute, which also is available from NTIS.

(instructional costs, administration, plant operation and maintenance, new construction, etc.) corresponding to school revenue sources. Such budgets do not show costs of specific program activities. For example, the costs of a specific reading program could not be determined from an administrative function budget: Teacher salaries would be included in one category, instruction materials in another, and building costs in a third. Thus, it was generally necessary for Education Turnkey to recast school budgets on a program basis.

Further, while much budgetary information is available on a district-wide basis, cost estimates of individual schools' activities frequently are not maintained. This is particularly true for indirect costs, such as plant operation and maintenance, which are not attributable to a single program. It was necessary in these cases to assign a portion of district-wide costs to individual schools. Because of the necessity of converting or imputing costs, data collection was most time-consuming. Assisting in cost data collection was one of the more difficult tasks for the districts' project directors. We are indebted to them and cooperating school personnel for their efforts to make this study possible.

METHODOLOGY

Education Turnkey Systems analyzed cost differentials between individual programs at specific sites and also differences in resource allocations among the programs. The results of these analyses are expressed in two ways. The first is in terms of local costs, which permits a comparison of each experimental program with its local control counterpart.

It was also desirable to compare the structural emphases of programs across sites such as the allocation of resources among professional and paraprofessional staff, instructional materials, and other forms of support. But in comparing programs carried out in different parts of the country, observed differences in resource allocation could occur because of both structural program differences and differences in the regional price differences. Therefore, average national prices for various program inputs were substituted and a "national average price" model of each program constructed. Both the structure of programs and their relative levels are analyzed in these terms. This was the second basis for stating costs.

Cost data were available from only 10 of the 18 districts because of difficulties encountered in either obtaining or validating data.[3]

---

[3] Cost analyses of the incentives-only sites were not undertaken because the cost of incentives provided to the experimental schools by OEO was a minor addition to the normal school expenditures.

However, at least one district was included for each of the six private firms. Within each district, cost models were constructed for each of the following cases:

Experimental	Elementary	Reading
Control	"	"
Experimental	Elementary	Math
Control	"	"
Experimental	Secondary	Reading
Control	"	"
Experimental	Secondary	Math
Control	"	"

The data were collected from the third grade program for the elementary models and the eighth grade for the secondary models. These were felt to be representative of two groups of grades included in the experiment.

Education Turnkey obtained the initial data from questionnaires filled out by representatives from the participating school districts and the private firms. This information was organized into a computer-based model known as Cost-Ed. Where data from schools were not directly available, as for the cost of some resources indirectly supporting instruction, assumed values were inserted by Education Turnkey. A preliminary output of the programmed information was then returned to the school districts and firms, which were asked to verify data, review any assumptions made and fill in missing information. After being certified as accurate by the appropriate parties, these data were used as the basis for the final models.

In the Cost-Ed model, a school day is viewed as being made up of a number of functions, some of which directly involve the student and some of which do not. The former category includes both subject matter instruction and nonacademic activities such as homeroom, physical education, recess, lunch and transportation. The latter category, not involving the student directly, includes administration by building principals and district-wide administration.

The total cost of a reading or math program calculated by the Cost-Ed model includes direct classroom-related costs of instruction in the subject plus a prorated share of the costs of all nonacademic functions listed in the preceding paragraph. The nonacademic functions are considered to be supportive of instruction, whether or not the student is directly involved in them. Each function is seen to consume one or more of following types of resources:

1) Staff
2) Facilities - including instructmental equipment
3) Curricular materials
4) Supplies and miscellaneous

In each model, the cost of a resource is governed by its hours in use. For example, the cost of a teacher, whose time was spent entirely in an experimental program (as was usually the case) would be prorated between various classes according to the time spent teaching in each. While only a portion of her salary would be allocable to a particular class or model, the entire amount would be a cost of the experimental program.

In the case of control programs, teachers frequently spent only part of the time teaching control students or subjects. Thus, only the portion of their salaries proportionate to the amount of time spent in such a function would be charged to the particular control models.

The cost of a resource in a function supporting instruction, such as a building principal's administration, was allocated to either reading and/or math in the same ratio as the length of instruction in the subject bore to the entire school day. In cases where only district-wide data were available, as for operations and building maintenance costs, an allocation of appropriate costs was first made to the school building and then to the subject, proportionate to its scheduling in the school day.

The following additional factors about costs in this study should be noted:

-- To reduce irrelevant differences in support costs between the experimental and control schools, one base was developed from the control school and used for both. Then the specific costs for operating the experimental reading and math programs were added to the base to obtain the experimental program costs, and the costs of the regular reading and math programs were added to obtain the control program costs. The control school used as a base for the model is the one in the district that contained the most students in the two grades studied. Therefore, in addition to program models' being a sample of grades, the control school base used may be a sample of control schools in the district.

-- The educational firms' costs are contained in the reading or math "instruction" function. Their administrative costs are not included since the analysis was designed in part to assist districts which might like to implement a performance contracting project without involving a private firm. The firms' administrative costs are shown separately, however. As noted earlier the cost of school functions; including administration, supporting the experimental and control instructional programs, are included in program costs.

-- The direct instructional costs of the programs are shown in this paper as well as the total costs. These permit a closer comparison of the instructional systems themselves. Instructional costs were omitted from the Education Turnkey report by agreement with OEO.

-- The hours of instruction in reading and math were frequently different for the control and experimental students. In such cases a difference in cost between the two programs is partly a function of different instructional times --i.e., costs of a program may appear lower because the instructional period was shorter. Distortion due to this factor is reduced when instructional costs only are studied.

-- The costs of programs with different length instructional periods can be adjusted by dividing them by equivalent periods of time; the effect of and qualifications about doing so are discussed in the "Findings" section. While illustrative, such computations should not be substituted for the conclusions in this study.

-- Various types of resources consumed are treated differently. Some, such as teachers' salaries and consumable materials, are measured by actual expenditures. Others, such as depreciation and maintenance costs, are prorated as a share of a building's expected life-time costs in order to exclude irrelevant factors such as age of a building or a recent, major purchase of non-instructional equipment. A resource purchased for the program but never used would not be recorded as a cost. But the cost of a movie projector, for example, bought before the program and used in it would be included as the proportion that the experimental usage represented of total life-time usage times the total life-time cost.

-- The costs in a model are not a function of who paid for them. In most programs, the salaries of experimental teachers were paid by the companies, but in some programs they were paid by the districts. The model includes them regardless of their source.

-- The cost of any subcontractor incentive payments to reward teachers, as for student performance is not included, since data were not available when the analysis was performed. The costs of any incentive rewards to students are included.

-- Costs occasioned by the experiment, such as data collection, are not included, primarily because they would not be incurred in an operational program.

The Cost-Ed model estimates normal costs that would be incurred in the ongoing operation of a performance contracting program. Since some start-up costs are excluded, however, it may underestimate the costs that would initially face a school district during its first year of a similar program.4/

4/ For a similar, but not identical, study of performance contract costs, see R-900/1HEW, Case Studies in Educational Performance Contracting: Conclusions and Implications, The Rand Corporation, December, 1971, (available from the Rand Corporation, 1700 Main St., Santa Monica, California 90405, for $3.00) Rand's evaluation was developed in terms of "comparable replication costs," which are most similar to the national average instructional costs in the Cost-Ed model. As with the Cost-Ed model, many developmental and administrative start-up costs are excluded. Both models measure the cost of resources that would be required for an in-house replication of a learning system, rather than the actual expenditures of the experimental program.

In specific aspects, however, the Rand and Cost-Ed models are different. For example, Rand assumed an average teacher salary of $12,000, compared to the Cost-Ed model's $9,025. Rand assumed that classroom space was available, and did not include a cost for it (except for remodelling); the Cost-Ed model includes the cost of all space used. Therefore, the specific cost levels reported in this study and by Rand should not be considered precisely comparable.

FINDINGS

Findings from the cost analysis are discussed first in terms of local prices and then in terms of the "national average price" model. Comparisons are made of the cost structure of the programs as well as their levels on the "national average price" basis. Finally, considerations affecting the comparability of program costs with different length instructional periods are discussed.

Experimental versus Control Comparisons - (local prices)

In 24 of the 40 site-grade-subject combinations studied, the experimental programs were at least 5 percent more costly than the controls. In six cases they were within 5 percent of the controls and in 10 the control programs were at least 5 percent more costly than the experimental. The standard of 5 percent has been adopted as a meaningful difference to limit the possibility of "program" differences being an artifact of sampling or statistical procedures used in the Cost-Ed model. In general, therefore, the experimental programs tended to be equal to or more costly than the controls. Table I shows the total cost per student year of the programs expressed in local prices.

The companies' on-site administrative costs per student year are shown separately in Table II. These range from $17.42 to $47.74 (local prices) per subject per grade. If added to each of the appropriate experimental cost figures in Table I, the values in Table

Table I

Total Cost Per Student Year (Local Prices)
and Class Hours Per Day

Elementary Programs

Site	Reading Costs (Hrs/Day)		Math Costs (Hrs/Day)	
	Experimental	Control	Experimental	Control
Grand Rapids	$191.80 (1.156)	157.44 (1.050)	$190.68 (1.156)	100.08 (.667)
Taft	188.19 (1.500)	190.83 (2.000)	$ 74.19 (.500)	$103.26 (1,083)
Hammond	$263.81 (.750)	271.41 (1.700)	259.94 (.750)	148.00 (.927)
Jacksonville	$244.12 (1.000)	142.83 (1.083)	225.32 (1.000)	98.90 (.750)
Athens	$172.40 (.920)	140.76 (1.100)	168.31 (.920)	106.60 (.833)
Selmer	$122.16 (.750)	169.38 (2.000)	117.64 (.750)	84.69 (1.000)
Dallas	$170.80 (1.000)	179.84 (1.546)	170.80 (1.000)	119.58 (1.028)
Portland	$216.76 (.917)	322.63 (1.917)	184.44 (.917)	112.22 (.667)
Seattle	$225.90 (.694)	306.75 (1.000)	224.79 (.722)	349.53 (1.520)
Fresno	$180.55 (1.000)	268.02 (1.500)	180.55 (1.000)	178.68 (1.000)

Table I (con't)

Total Cost Per Student Year (Local Prices)
and Class Hours Per Day

Secondary Programs

Site	Reading Costs (Hrs/Day)		Math Costs (Hrs/Day)	
	Experimental	Control	Experimental	Control
Grand Rapids	167.94 (.806)	180.66 (.917)	157.88 (.806)	180.66 (.917)
Taft	135.92 (.917)	156.92 (.917)	140.44 (.917)	143.07 (.917)
Hammond	208.18 (1.000)	128.54 (.717)	196.17 (1.000)	128.54 (.717)
Jacksonville	201.25 (.830)	133.27 (.833)	225.81 (.840)	133.27 (.833)
Athens	144.30 (.750)	142.44 (.833)	141.00 (.750)	142.44 (.833)
Selmer	184.93 (.750)	95.77 (1.000)	180.41 (.750)	95.77 (1.000)
Dallas	176.76 (1.000)	131.92 (.917)	179.85 (1.000)	131.92 (.917)
Portland	204.42 (.917)	137.66 (.726)	196.95 (.917)	137.66 (.726)
Seattle	243.55 (.917)	182.75 (.889)	233.50 (.917)	190.16 (.889)
Fresno	169.65 (.786)	137.67 (.7500)	169.65 (.786)	129.76 (.750)

Table II

Company Project Administration Costs by Site

Subcontractor	Site	Project Administration Cost Per Student-Year For Each Grade Level and Subject - (Local Prices)
Alpha Learning Systems	Grand Rapids	$28.57
	Taft	17.42
Learning Foundations	Hammond	23.95
	Jacksonville	20.77
Plan Education Centers	Athens	36.90
	Selmer	36.78
Quality Educational Development	Dallas	27.20
Singer-Graflex	Portland	43.72
	Seattle	46.19
Westinghouse Learning Corp.	Fresno	46.20 (Elem.)
		49.74 (Sec.)

II would alter the relationship of experimental to control program costs in some instances. In only three of ten cases in which the experimental program was at least five percent less costly than the control, less than a five percent difference would remain after addition of the administrative costs. In each of the six cases where the experimental program was less than five percent different from the control, the addition of values from Table II would make the experimental program at least five percent more costly than the control.

While incurred during the experimental year, these administrative costs were purposely not included in Table I. As stated, such costs (probably somewhat lower) would normally be incurred if these programs became operational. To reflect these costs, a portion of the school principal's salary and district-wide administration was allocated to the costs of the experimental programs. It is likely that some company administrative expenses would continue to be incurred if the experimental programs were replicated on an operational basis; however, the information-gathering requirements made upon all parties in the experiment, including the companies, increased these costs sustantially over those to be expected in normal program operations.

Analysis of "National Price" Costs

As noted earlier, substitutions of "national" average values for actual local costs of significant program parameters were made to account for regional differences in the costs of program inputs.

Any economic factor which was part of the structure of a program was not altered. For example, structural factors such as student staff ratios, time utilization and classroom square footage per/student in actual programs were all considered program specific and not altered. The costs for these factors, however, were not considered program specific, but related to the nature of the local economy. Local costs for these factors, were deleted and replaced with national average values. The main factors for which national average values were substituted were: salary and fringe rates of professionals and paraprofessionals contained in direct instructional costs and building acquistion, operations and maintenance included in supportive costs.

In principle, in the "national" average price model apparent differences in the allocation of resources are due solely to differences in program structure. While program costs are thereby made comparable across sites, such a substitution of national prices could affect the relative local costs of a control and experimental program in a specific site where the two programs made much different use of a factor with widely disparate national and local price values. The overall relationship of experimental to control program costs discussed in this paper is similar on the national average price and local bases; therefore this is not a significant problem.

Tables III-VI show the relationships between the pairs of experimental and control programs on the national average price basis to be similar to those expressed in terms of local prices. Using the national average price figures, 25 experimental programs were at least five percent higher than their controls and thirteen were at least five percent lower. The difference in cost between the remaining two was less than five percent.

Comparing costs across sites on the national overage price basis, the experimental costs for elementary reading tend to be lower overall than the controls, while other experimental costs are higher. For elementary reading, the median experimental cost for the ten sites is $217 and median control cost is $254; for elementary math, the median experimental cost is $200 and control $136; for secondary reading, the median experimental cost is $218 and control $169; and for secondary math, the median experimental cost is $216 and control $169.

It is desirable to compare the direct costs of instruction for the experimental and control programs as well as their total costs. To recall, the total cost figures include an allowance for functions supporting instruction. These constitute approximately 35 to 50 percent of total control program costs and 20 to 40 percent of experimental program costs. (The absoluate levels of supportive costs are more similar, but those in experimental programs constitute a lower percentage of the higher total costs.)

Table III
Total Cost Per Student-Year (National Average Prices)

Elementary Reading

Rank	Experimental District	Cost	Rank	Control District	Cost
1	Selmer	$147.70	1	Athens	$150.32
2	Dallas	186.47	2	Jacksonville	175.53
3	Athens	190.84	3	Grand Rapids	186.57
4	Fresno	215.52	4	Dallas	216.63
5	Seattle	215.79	5	Seattle	252.35
6	Grand Rapids	217.29	6	Selmer	255.76
7	Hammond	252.04	7	Hammond	274.15
8	Portland	263.01	8	Fresno	286.95
9	Jacksonville	270.25	9	Taft (Sinton)	300.85
10	Taft	280.52	10	Portland	349.80

Table IV
Total Cost Per Student-Year (National Average Prices)[1/]
(See Footnote)

Elementary Math

Rank	Experimental District	Cost	Rank	Control District	Cost
1	Taft	$104.94	1	Athens	$113.83
2	Selmer	143.18	2	Grand Rapids	118.52
3	Dallas	186.47	3	Portland	121.61
4	Athens	186.76	4	Jacksonville	122.24
5	Seattle	214.16[1/]	5	Selmer	127.88
6	Portland	212.76	6	Dallas	143.92
7	Fresno	215.52	7	Hammond	149.50
8	Grand Rapids	216.17	8	Taft (Sinton)	162.77
9	Hammond	248.20	9	Fresno	191.30
10	Jacksonville	251.76	10	Seattle	288.03

[1/] Corrected from Education Turnkey final report.

Table V
Total Cost Per Student-Year (National Average Prices)

Secondary Reading

Rank	Experimental District	Cost	Rank	Control District	Cost
1	Athens	$176.57	1	Hammond	$148.98
2	Grand Rapids	182.66	2	Portland	153.52
3	Taft	186.68	3	Selmer	158.84
4	Fresno	201.68	4	Fresno	159.01
5	Dallas	212.79	5	Jacksonville	163.43
6	Jacksonville	223.55	6	Dallas	173.96
7	Hammond	227.54	7	Seattle	175.93
8	Selmer	231.59	8	Grand Rapids	178.16
9	Portland	253.97	9	Athens	188.85
10	Seattle	262.68	10	Taft (Sinton)	223.60

Table VI
Total Cost Per Student-Year (National Average Prices)

Secondary Math

Rank	Experimental District	Cost	Rank	Control District	Cost
1	Grand Rapids	$171.65	1	Hammond	$148.98
2	Athens	173.27	2	Fresno	151.43
3	Taft	192.71	3	Portland	153.52
4	Fresno	201.68	4	Selmer	158.84
5	Hammond	215.57	5	Jacksonville	163.43
6	Dallas	216.41	6	Dallas	173.96
7	Selmer	227.07	7	Grand Rapids	178.16
8	Portland	245.88	8	Seattle	182.48
9	Jacksonville	248.85	9	Athens	188.85
10	Seattle	254.09	10	Taft (Sinton)	206.36

When the supportive costs are set aside any cost differences due to variation in instructional time not related to the programs themselves are eliminated and a better comparison may be made between costs of direct instruction. Any differential use of instructional resources by the experimental and control programs, such as staff, equipment and materials and also any cost differences directly related to different length instructional periods, is best illustrated on this basis. Inasmuch as all company costs are contained in the instruction function, a closer comparison of company and school costs is also facilitated.[4] Instructional costs for experimental and control programs are portrayed separately in Tables VII-X.

Tables VII-X show the direct instructional costs of experimental programs to be higher to a slightly greater extent than was the case for total costs. Twenty-eight experimental programs have instructional costs at least 5 percent higher than the controls and eight at least 5 percent less costly; the remaining four are within 5 percent of the controls. Analysis of these figures also indicates the experimental programs differed significantly from controls in the pattern of costs incurred for instructional resources.

---

[4] Educational company costs constitute the bulk of the instruction function in experimental programs, with the exception of a prorated charge for classroom acquisition, operation, and maintenance borne by the schools. Also, in Grand Rapids, Hartfor, and Taft, the cost of professional teachers in the experimental programs was paid by the schools.

Table VII
Programs Ranked by Instructional Cost Per Student-Year (National Average Prices) Showing Distribution of Costs (See Footnotes)

Elementary Reading

Rank	Site	Instructional Cost Per Student-Year 1/	Percentage of Instructional Cost: Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Control Programs									
1	Athens	$ 77.48	70.9	0	70.9	23.0	2.3	1.1	2.7
2	Jacksonville	99.59	66.9	0	66.9	26.7	1.0	0.9	4.5
3	Grand Rapids	104.72	64.4	11.6	76.0	17.7	1.9	2.8	1.6
4	Dallas	118.88	80.9	0	80.9	15.3	1.7	0.8	1.3
5	Seattle	162.81	77.7	0	77.7	12.1	1.0	3.8	5.4
6	Selmer	145.59	78.8	0	78.8	18.4	0.5	1.0	1.3
7	Hammond	160.44	73.8	0	73.8	23.6	0.9	0.7	1.0
8	Fresno	158.88	75.4	0	75.4	19.4	0.9	1.4	2.9
9	Taft (Sinton)	171.90	80.0	0	80.0	17.9	0.4	0.6	1.1
10	Portland	211.86	74.3	0	74.3	19.9	2.3	2.3	1.2
Experimental Programs									
1	Selmer	$106.42	46.1	15.4	61.5	19.2	0.8	17.6	0.9
2	Dallas	123.23	53.0	19.0	72.0	13.3	5.3	6.1	3.3
3	Athens	129.91	53.2	19.7	72.9	10.9	1.5	14.7	0
4	Fresno	130.13	12.6	41.7	54.3	10.4	8.6	3.6	23.1
5	Seattle	153.60	52.2	9.5	61.7	8.4	4.1	23.3	2.5

Table VII (Cont.)
Programs Ranked by Instructional Cost Per Student-Year (National Average Prices)
Showing Distribution of Costs (See Footnotes)

Elementary Reading

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost						
			Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Experimental Programs									
6	Grand Rapids	$127.18	51.5	15.0	66.5	20.3	11.9[2]	0[2]	1.3
7	Hammond	201.87	0	59.6	59.6	7.0	2.3	24.9	6.2
8	Portland	197.01	66.4	0	66.4	13.2	5.5	13.3	1.6
9	Jacksonville	139.20	0	66.1	66.1	7.9	2.1	17.6	6.3
10	Taft	183.81	49.5	32.0	81.5	9.4	0	8.2	0.9

---

[1] Similar tables (XIII-XVI) were contained in the Education Turnkey final report in which total cost per year were shown together with the percentage distribution of instructional costs. The percentages in that report apply to the instructional costs reported in this paper, Tables VII-X.

[2] Corrected from Education Turnkey final report.

Table VIII
Programs Ranked by Instructional Cost Per Student-Year (National Average Prices) Showing Distribution of Costs

Elementary Mathematics

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost						
			Teachers %	Paraprofessional %	Total Staff %	Class-room %	Instructional Equipment %	Books and Audiovisual %	Other %
Control Programs									
1	Athens	$ 58.67	70.9	0	70.9	23.0	2.3	1.1	2.7
2	Grand Rapids	66.52	64.4	11.6	76.0	17.7	1.9	2.8	1.6
3	Portland	73.68	74.3	0	74.3	19.9	2.3	2.3	1.2
4	Jacksonville	68.97	66.9	0	66.9	26.7	1.0	0.9	4.5
5	Selmer	72.79	78.8	0	78.8	18.4	0.5	1.0	1.3
6	Dallas	78.93	80.9	0	80.9	15.3	1.7	0.8	1.3
7	Hammond	87.49	73.8	0	73.8	23.6	0.9	0.7	1.0
8	Taft (Sinton)	93.09	80.0	0	80.0	17.9	0.4	0.6	1.1
9	Fresno	105.92	75.4	0	75.4	19.4	0.9	1.4	2.9
10	Seattle	151.27	63.6	0	63.6	19.8	1.6	6.3	8.7
Experimental Programs									
1	Taft	$ 72.75	54.6	16.1	70.7	7.9	0	20.7	0.7
2	Selmer	101.89	48.1	16.1	64.2	20.1	0.9	14.6	0.2
3	Dallas	123.23	53.0	19.0	72.0	13.3	5.3	6.1	3.3
4	Athens	125.84	54.9	20.3	75.2	11.3	1.5	11.9	0.1
5	Seattle	149.46	55.8[3/]	10.2[3/]	66.0[3/]	9.0[3/]	4.3[3/]	18.3[3/]	2.5[3/]

Table VIII (Cont.)
Programs Ranked by Instructional Cost Per Student-Year (National Average Prices) Showing Distribution of Costs

Elementary Mathematics

			Percentage of Instructional Cost						
Rank	Site	Instructional Cost Per Student-Year	Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Experimental Programs									
6	Portland	$146.76	44.6	14.3	58.9	17.8	1.4	19.8	2.1
7	Fresno	130.13	12.6	41.7	54.3	10.4	8.6	3.6	23.1
8	Grand Rapids	126.06	52.0	15.1	67.1	20.5	0	12.0	0.4
9	Hammond	198.02	0	60.7	60.7	7.2	1.4	24.4	6.3
10	Jacksonville	180.71	0	72.9	72.9	8.7	1.2	10.3	6.9

---

[3] Corrected from Education Turnkey final report.

Table IX

Programs Ranked by Instructional Cost Per Student-Year (National Average Prices) Showing Distribution of Costs

Secondary Reading

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost						
			Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Control Programs									
1	Hammond	$ 82.45	67.3	0	67.3	30.0	0.8	0.7	1.3
2	Portland	75.00	68.4	0	68.4	23.4	1.6	3.1	3.5
3	Selmer	79.00	73.9	0	73.9	23.1	0.6	1.0	1.4
4	Fresno	78.26	70.8	0	70.8	20.1	1.3	5.0	2.8
5	Jacksonville	81.31	73.1	0	73.1	19.9	1.4	1.0	4.6
6	Dallas	98.33	77.6	0	77.6	19.6	1.3	0.5	1.0
7	Seattle	101.29	70.3	0	70.3	14.9	0.8	6.2	7.8
8	Grand Rapids	104.42	72.4	0	72.4	20.4	1.5	3.9	1.8
9	Athens	106.73	68.3	0	68.3	26.9	2.3	0.7	1.8
10	Taft (Sinton)	147.56	70.1	0	70.1	27.8	0.5	0.4	1.2
Experimental Programs									
1	Athens	$102.66	48.8	18.0	66.9	13.7	4.3	15.2	0
2	Grand Rapids	117.88	43.9	32.7	76.6	13.8	0.3	5.6	3.7
3	Taft	110.63	58.1	18.0	76.1	13.8	0.4	6.0	3.9
4	Fresno	117.05	14.7	32.0	46.7	14.0	9.6	4.0	25.8

Table IX (con't)

Programs Ranked by Instructional Cost Per Student-Year (National Average Prices)
Showing Distribution of Costs

Secondary Reading

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost						
			Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Experimental Programs									
5	Dallas	130.30	52.9	18.0	70.9	14.6	4.9	5.8	3.8
6	Jacksonville	141.43	0	55.7	55.7	11.4	5.9	18.2	8.8
7	Hammond	134.68	0	42.9	42.9	13.3	10.9	23.7	9.3
8	Selmer	171.71	60.1	19.1	79.2	8.2	1.3	10.9	0.6
9	Portland	154.78	66.8	0	66.8	14.3	7.0	9.9	2.0
10	Seattle	185.65	46.7	6.1	52.8	30.6	3.4	11.1	2.1

Table X

Programs Ranked by Instructional Cost Per Student-Year (National Average Prices)
Showing Distribution of Costs

Secondary Mathematics

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost: Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Control Programs									
1	Hammond	$ 82.45	67.3	0	67.3	30.0	0.8	0.7	1.3
2	Fresno	70.68	71.3	0	71.3	20.4	1.4	3.8	3.1
3	Portland	75.00	68.4	0	68.4	23.4	1.6	3.1	3.5
4	Selmer	79.00	73.9	0	73.9	23.1	0.6	1.0	1.4
5	Jacksonville	81.31	73.1	0	73.1	19.9	1.4	1.0	4.6
6	Dallas	98.33	77.6	0	77.6	19.6	1.3	0.5	1.0
7	Grand Rapids	104.42	72.4	0	72.4	20.4	1.5	3.9	1.8
8	Seattle	107.83	71.1	0	71.1	15.0	0.7	5.8	7.4
9	Athens	106.73	68.3	0	68.3	26.9	2.3	0.7	1.8
10	Taft (Sinton)	130.32	66.1	0	66.1	31.5	0.5	0.5	1.4
Experimental Programs									
1	Grand Rapids	$106.87	48.4	25.8	74.2	15.2	0.3	6.2	4.1
2	Athens	99.36	50.4	18.6	69.0	14.2	4.4	12.2	0.2
3	Taft	116.66	63.0	14.3	77.3	12.9	0.4	5.7	3.7
4	Fresno	117.05	14.7	32.0	46.7	13.9	9.6	4.0	25.8

Table X (con't)

Programs Ranked by Instructional Cost Per Student-Year (National Average Prices) Showing Distribution of Costs

Secondary Mathematics

Rank	Site	Instructional Cost Per Student-Year	Percentage of Instructional Cost						
			Teachers %	Paraprofessional %	Total Staff %	Classroom %	Instructional Equipment %	Books and Audiovisual %	Other %
Experimental Programs									
5	Hammond	122.71	0	47.1	47.1	14.5	8.4	19.8	10.2
6	Dallas	133.92	51.5	17.4	68.9	16.9	4.8	5.6	3.8
7	Selmer	167.19	61.7	19.6	81.3	8.4	1.3	8.9	0.1
8	Portland	146.70	70.5	0	70.5	15.1	1.5	10.8	2.1
9	Jacksonville	166.73	0	47.8	47.8	9.7	21.6	13.4	7.5
10	Seattle	177.06	48.4	4.8	53.2	32.1	3.6	9.0	2.1

All experimental programs incurred costs for paraprofessional staffing, while only Grand Rapids did so among the controls. Experimental programs in Hammond, Jacksonville, and Fresno incurred little or no costs for professional teachers, although costs for paraprofessional staffing in the first two programs were equal to or higher than costs for professional staffing in their control counterparts. On an aggregate basis, the costs of experimental staffing, including paraprofessionals, is 27 percent higher than control. When paraprofessionals are excluded, the cost of professional teachers in experimental programs is 72 percent that of controls. However, when the three experimental programs which relied substanitally on paraprofessionals are excluded, the cost of professional teachers in the remaining experimental programs is equal to their controls.

Experimental programs incurred significantly higher costs for books and audiovisual software and to a lesser extent for instructional equipment such as teaching machines. In the aggregate, the cost of instructional hardware in the experimental programs is four times higher than in the control programs and the experimental cost for books and audiovisual software is 10 times higher than for controls.

Experimental programs also differ among themselves and between elementary and secondary practices in patterns of costs incurred. The Grand Rapids and Taft programs incurred almost no costs for instructional hardware, while most others show moderate to heavy costs for such equipment. The Jacksonville, Hammond, and Athens programs show higher costs for hardware for the secondary than for the elementary grades, while these costs are more evenly distributed among grades in the other experimental programs.

The highest experimental cost level for books and audiovisual curricular materials among the elementary grades was in Hammond, where it approximated $50 per student. In these programs, materials were 25 percent of total instructional costs. The Dallas programs incurred costs of $7.52, the lowest for such materials among the elementary grades. This represented only 6 percent of total instructional costs. By comparison, cost for curricular materials in most control programs was less than 3 percent of instructional costs.[5/]

The eight experimental programs whose instructional costs were 5 percent less expensive than their controls on the national average price basis do not share any consistent pattern of resource utilization.

They generally offset higher costs in the area of instructional equipment or curricula with lower staffing costs, or the reverse; however they are not consistently low cost in either technology or staffing. One frequently contributing factor to their lower cost is shorter instructional time.

[5/] Estimates for costs of instructional equipment and curricular materials in control programs are based on prorations of district-wide costs; therefore comparison with control programs should be considered only approximate.

The Effect of Instructional Time on Cost Models

The instructional time available per day for a subject differed frequently between experimental and control schools. In 18 instances studied, the control program was longer than the experimental program. In eight there was less than a 5 percent difference and in 14 the experimental program was longer than the control. The trend toward shorter experimental programs was most pronounced at the elementary level, and particularly for reading, where the experimental program was shorter in eight of 10 instances.[6/] As noted, cost estimates are in part a function of a daily instructional time; therefore any difference between experimental and control costs in programs with different length instructional periods is due in part to the time difference only.

The cost per year figures could be adjusted to account for these differences by dividing each by the proportion its actual daily instructional time in minutes bears to one hour.

[6/] Precise reporting of time spent in reading instruction in the elementary control schools was difficult. Specific instruction in reading skills, silent reading periods, story telling, or language arts all occurred but only time spent in specific instruction was to be reported. In the experimental programs, where children were scheduled into learning centers for fixed periods of instruction, reporting was easier. It is possible that the instructional time for some control programs is overstated. The problem existed only to a minor extent for elementary math control programs and not in the secondary grades for which scheduled periods were reported.

The resulting figures would express costs in terms of an equivalent yearly rate; programs shorter than one hour per day would assume a higher rate and those longer than one hour a lower rate.

The reader is cautioned against drawing unqualified conclusions about "adjusted" costs, however. In the control programs all instructional costs were based on proration of existing school/school resources. In the experimental programs, however, there is not necessarily a 1:1 relationship between adjusted time and adjusted costs. While the cost of each firm's resources has been allocated among individual experimental classes in a district, based on their relative length of time, the gross amount of a firm's investment in a district does not bear a determinate relationship to the length of its instructional periods. It is likely, for example, that if a program were actually extended, increased staffing costs would be incurred, but not necessarily proportionate to the increase in time; other offsetting changes might occur. Also, the instructional equipment and curricula are not likely to be increased by longer exposure in classrooms. Therefore, while adjusted figures suggest further differences between experimental and control costs, they illustrate the rate of costs actually incurred and not, in the case of the experimental programs, costs that would necessarily be incurred if their length was actually altered in the manner assumed by the adjustment. Evaluation findings are, of course, based on actual program characteristics and costs; therefore any cost-effectiveness studies would be restricted to actual program data.

Instructional costs have been presented separately in part to eliminate time-related cost differences attributable to support functions. If the instructional costs are adjusted for time factors only two of the eight experimental programs which had instructional costs at least 5 percent lower than the controls remain lower. (Taft - secondary reading and math). This indicates a generally higher rate of cost incurrence in the experimental programs, were pronounced than is evident when examining cost per year figures. Due to the reasons cited, however, this finding must be regarded as theoretical.

SUMMARY

On the whole, then, the performance contracting programs were found to be generally more costly than control programs in the experiment.

To a large extent, the individual performance contract programs that were less costly had shorter instructional periods. While there may be a significant potential for reducing costs through shortening programs, this saving will not clearly be achieved if the short program is substituted for a longer one in the contest of a normal school day.

Direct instructional as well as total costs have been described in this paper. Instructional costs better illustrate differences between experimental and control programs and reduce the extent of differentation introduced by time factors alone. However, even when instructional costs are considered alone, the experimental programs are found to be generally more expensive than the control programs.

Staffing costs in performance contract programs were as high or higher than control programs, even though in some cases paraprofessionals were substituted for professionals. The performance contract programs also incurred higher costs for educational equipment and materials.

In summary, the performance contract programs in this experiment are not less expensive alternatives to present educational programs.

Chapter V

PROJECT MANAGERS' STATEMENT

INTRODUCTION

The primary basis for evaluation of the OEO performance incentive contract experiment must be measured student gains in reading and mathematics skills. There can be little debate on this matter; it is and has been clearly understood by all parties involved. The Office of Economic Opportunity has assumed the task of providing the basic evaluation design to yield the necessary statistical analysis and interpretation of the test scores. At this writing the project directors have not been provided access to the test scores and, consequently, are in no position to make comment in this most critical substantive area.[1]

The purpose of this chapter is to provide a means through which the project directors might collectively relate their perceptions as well as express some concerns on procedural matters and interrelationships in the OEO projects. This chapter provides a look at the project and some of its problems as seen from the local level.

Whether or not the statistical evaluation supports or discourages the concept of performance contracting, it is obvious that private enterprise will continue in some relationship with the nation's public schools, and that the U. S. Government will continue to encourage educational research and program development. We feel that the year's experiment has provided some significant experience in the relationships among the Government, private business and the public schools and that a frank discussion may serve to help to avoid future pitfalls as such relationships develop.

---

[1] This statement was prepared in January, 1972. The project directors have since been provided with complete evaluation test results.

Although the critique which this chapter represents may, by its nature, appear negative, it is the combined feeling of the project directors that this must be tempered in the mind of the reader by the fact that OEO had the fortitude to take the bold step in sponsoring this project in full realization of its inevitably controversial nature. The project directors also feel that mention must be made of the fact that representatives from many private businesses with whom we worked were not, in our opinion, motivated solely by potential profit but were sincerely trying to find ways to solve some of our most difficult, complex and frustrating problems in education.

Care has been taken in this chapter to generalize our comments because for the most part they represent opinions, perceptions and individual experiences, otherwise subcontractors could be unfairly damaged by a stress on specifics. We also feel that the experimental nature of the program created some procedural problems that could not have been reasonably anticipated.

It must be said, however, that an amazing degree of consensus exists in the perceptions by most project directors in regard to strengths and weaknesses of the program.

Five major areas of concern are covered. They are (1) project start-up constraints, (2) program implementation, (3) subcontractor programs, (4) critique of management subcontractor and (5) critique of test and analysis contractor. These are followed by some major conclusions and recommendations.

PROJECT START-UP CONSTRAINTS

Constraints arose in launching the project.

(1) Late pre-planning of project with the various agencies involved.

(2) Lack of Local Education Agency i.e., local district personnel, involvement in the initial stages of project.

(3) Late selection of testing and analysis contractor

(4) Inadequate pre-service training of local staffs

Most districts were approached by OEO in late May and asked to send representatives to Washington, D. C. in mid June. All negotiations on the original prime contracts took place in one day. Although this was an expedient manner of handling negotiations, many school districts were forced to make a "go-no-go" decision without full knowledge of all the implications involved if not negative attitude.

Since negotiations took place during the summer, most school personnel did not know they would be involved until school opened in the fall. This short lead time in initiating the project caused many teachers and local unit administrators to view the project with apprehension.

The August selection of the testing and analysis contractor presented major problems in setting up pre-test planning and administration.

The pre-service training of project staffs was hampered by the unfamiliarity of some subcontractor's project administrators with the instructional program. Often, there was an absence of most materials and equipment to be utilized during the training workshop. Since most subcontractors were using commercially supplied materials, part of the problem could be charged to the procurement process in acquiring them from publishers and distributors.

All of these problems center around a lack of sufficient time. There is and has been conjecture on the part of many, particularly those agencies who fund programs of this nature, about the trade off between a long lead time to do sufficient planning and a short lead time which forces the "systems" to not become involved in the processes involved. It is the consensus of the project directors involved that the former would have been the preferred procedure.

PROGRAM IMPLEMENTATION

As project directors we feel that one of the major factors that served as a deterrent to subcontractors being prepared to implement their programs at the start of school at various sites was the ill timing in the hiring of project administrators.2/ Even though some project administrators were employed prior to the negotiating and finalizing of subcontracts, others were not brought aboard until a few days before the start of school or after the project had been implemented. This delay served as the main cause for the inadequate pre-service training given to subcontractor personnel. Depending upon the subcontractor and site this training time ranged from three days to two weeks.

Even in sites in which a greater number of pre-school training days was available, the effectiveness of this training was severely hampered by the lack of subcontractor materials being available for demonstration and practice.

After programs had been implemented and operating, some subcontractors found it necessary to modify their programs. This modification almost always meant a change in materials and supplies utilized. At the programs' conclusions, almost all subcontractors were using similar core programs. This modification, by no means, discounts the fact that subcontractors did bring into the program, previously used successful systems.

2/ Project administrators were the on-site representatives of the education technology companies.

It does, however, point out that while a program might be most effective with one particular student population, it might be relatively ineffective with others.

SUBCONTRACTORS PROGRAMS

The contingency management system used by five of the six subcontractors was innovative. Material incentives have been used before, but not on such a large scale for behavior modification or in such various and unique ways. Other innovations included the materials management approach which individualized instruction on a mass basis. This has many implications for the public schools such as: mass remedial education approaches, a method of breaking the grade-level lock-step system, more efficient use of paraprofessionals, flexibility in curriculum, introduction of teaching machines into the school arena, measuring of teaching productivity through student standarized testing, guaranteeing of student performance as a condition of payment, utilization of the systems approach in education, cost-effectiveness and internal school organizational reform.

A wide variety of software and hardware was introduced into the school systems. Subcontractor programs varied in the use of these. Some subcontractors used all software, all hardware or a combination of both.

Although a wide variety of hardware was used in some projects, it was not of the type that was new to school systems. The blend, mix and management of materials making up the instructional strategy were the things that were new to school systems.

Student incentives; rewards for scholastic achievement and/or behavior modification, were successful when properly handled. Student incentives had an adverse or no effect when improperly administered i.e., when prizes to be awarded by a few subcontractors were not readily available for presentation to students. Inappropriate gift selection also created a problem.

One contractor's token economy and reinforcing events room appeared to be highly successful because rewards were immediately available to students.

Staff incentives were used in a few projects. These were administered differently at each site due to local teacher union contracts or local school district situations. Their impact has not been measured at this writing.

Most subcontractors used a diagnostic system of determining pupils weaknesses then prescribed a remedial program of individualized curriculum for the students.

Project directors questioned the inclusion of first grade students in a remedial program of this type. Readiness programs were lacking and had to be designed by the staffs. Inclusion of first graders also created complications in achievement testing.

As the programs progressed it appeared that each company did not have an individual or unique curriculum approach. This is supported by the fact that many companies used the same or similar core instructional materials. This may indicate the lack of adequate published materials for remedial purposes.

A charge has been leveled that the hardware and the procedures used in performance contracting could be dehumanizing for students. It is the consensus of the project directors that the opposite was true because machines provided opportunities for more individual attention by creating additional learning activities for students and freeing staffs to interact on an individual basis with students. As the adult-pupil ratio was lowered and staffs freed to assist individual students, student skills improved and self-concepts were enhanced.

Each subcontractor was responsible for the teaching of basic skills in reading and arithmetic for experimental pupils. There was concern from site to site among regular classroom teachers as to where basic skills in reading left off and Language Arts began. This is generally not debated in non-contract elementary school situations since the same classroom teacher is responsible for the total curriculum. The separation of instructional functions requires team co-ordination of activities.

The handling of staff personnel varied from site to site and seemed to reflect the background and competence of the subcontractors project administration. In future performance contracts, all personnel should be hired and evaluated by the school district and education technology company. Among the competencies needed most in a project administrator is knowledge of and the ability to work within the framework of a public school system.

Additional lead time would have allowed for a clearer definition of roles of all personnel. The role of the project director varied from site to site as to the degree of his involvement in the program. The local administrative structure was at times in a quandary as to how to react to the project administrator, a new educational leader on the scene. On some sites, teaching staffs were confused as to who was the educational leader - the principal who has traditionally held this role or the on-site consultant who dictates (under contract) the curriculum or the project director.

Subcontractors appear to have difficulty in providing the data required by the management support contractor. Contributing factors were geographical location of buildings, lack of orientation to required documentation systems or recognition of the importance of data collection.

CRITIQUE OF MANAGEMENT SUPPORT CONTRACTOR

A management support group (MSG) functioned in the project as a liaison between the OEO and local education agencies (LEA). It was the feeling among most project directors that although MSG was given the responsibility of providing management support to the LEA, in fact they did not have the authority to make decisions in the project operations. This was probably due to the lack of definition of roles and authority assignments between OEO, MSG, and LEA.

Some facets of the Cost-Ed model developed by MSG for the project sites, are being questioned, due to a lack of complete understanding.

(1) How valid is the model? The project directors are concerned with the manner in which data was collected and substantiated at the local level. The thoroughness with which the model was built varied from site to site. This variance was probably due in part to a combination of the cooperation encountered with the local education personnel and the availability of data. In addition, there were sites at which there was no follow-up by MSG to the initial visits. Because of this lack of uniformity and perhaps understanding, the use of the Cost-Ed model as an evaluative instrument in comparing programs and program costs is felt to be of questionable value.

(2) The exclusion of the subcontractor's present administration cost in the cost comparisons of the various sites was unfortunate.

It is felt that local districts cost comparisons with subcontractor costs are open to question as a result of this omission.

(3) The analysis of the data comparing the total per-pupil cost of the experimental and control programs is suspect. Specific concern has been voiced over the method by which the cost of the classroom area and instructional time at the control schools resulted in off-setting the high cost of learning equipment and materials used at the experimental sites.

Project directors unanimously feel that the concept of the model is commendable. However, they are concerned about its present usefulness as a tool for program budgeting and instructional systems design and/or redesign.

It is unfortunate that MSG did not follow up their Cost-Ed model with an adequate explanation of its use to the LEA. The fact that it was sent to sites in late August prevented most school systems from utilizing the information for program planning and budgeting for the current school year.

Project directors question the curriculum audit conducted at some project sites. The interpretations of the design varied from site to site as well as the manner in which the audit was conducted. This difference was probably due to the personnel conducting the audit. There are charges that some audits consisted of walking in and out of classrooms and in some cases not visiting classrooms at all.

There are also concerns about the usefulness of the voluminous amount of documentation required in the project. To date, sites have not received analysis or results of findings of some data. Data collection often tended to alienate local district personnel as well as project personnel. However, project directors support the data collection concept as a valid contractual agreement. The concern lies with the amount and lack of feedback results.

Project directors feel that the concept of management support is valid and once refined, could be a most valuable tool as a liaison between LEA's and subcontractors and other agencies negotiating in the school arena.

CRITIQUE OF TEST AND ANALYSIS CONTRACTOR

In theory the experimental project design which had a test and analysis contractor (TAC) evaluating the entire experiment was good, but in practicality the following shortcomings were apparent:

(1) A lack of adequate pre-planning time for determining pretest sites, selection of students, adequate selection and training of testers, and test booklet preparations created hardships for local districts. The demands made by TAC were not made known to the LEA's soon enough and TAC's representative arrived at sites without a clear picture of his role and responsibilities regarding the testing program.

(2) At some sites the personnel utilized by TAC were quite competent; however, at other sites, the lack of experience in planning and arranging for mass student testing proved a real liability. In addition, some school districts also lacked the ability to handle the mass testing of students.

(3) TAC did not supply pre-test print out information to the sites until late fall. This made identification and proper placement of students who had actually been pretested a frustrating task. In some cases this meant administering pretests to some students as late as January of 1971, as a make-up procedure.

(4) Interim Performance Objectives (IPO) tests designed to approximate criterion - referenced tests were not utilized as such. Project directors feel that these tests were used as a routine to stimulate cash flow for the subcontractors.

(5) Post-test conditions were reported by project directors as being improved over the pretesting. This was due mainly to the fact that there was a sufficient amount of lead time, plus the experience gained from the pre-tests both by the LEA's and TAC.

(6) Retention testing as a final TAC responsibility cannot be evaluated in this report as it is being completed at this time.

SUMMARY AND RECOMMENDATIONS

The project directors, on the basis of the year's experience, feel a need to make recommendations: 1) the experiment as a unique procedural entity, and 2) performance contracting in general. The reader must keep in mind that project directors are attempting to provide guidelines, and the implications of their comments should not be construed as other than their subjective reactions; exceptions were found in each of the areas, therefore, recommendations must be viewed in light of the applicable situation.

I. The Experiment as a Unique Entity

1. Lines of communication in an experiment of such magnitude must be an area of major concern and effort. A concerted attempt was made to create and maintain openness of communication among all parties, but the complexity of the program led to numerous instances of confusion and frustration that could have been avoided.
2. Definition of roles must be provided at all levels. Once again, an attempt was made to define functions of all personnel, but it is clear that perceptions vary.
3. It is strongly recommended in an experiment such as this that project directors or district leaders be brought together on a regular basis. The initial regional conferences in August of 1970 were helpful, but was directed to people who had little opportunity to be in a position to understand what was about to happen.

No further opportunity was provided, except on an individual visitation basis, for project directors to profit by their collective experience until the program was completed. When a conference was called at the conclusion of the project, project directors were amazed to find the universality of the problems they had encountered. It was unfortunate that such problems and possible solutions could not have been shared while the program was in progress.

4. It is recommended that maximum effort be made by outside agencies to understand and to function within the structures placed upon local districts by states and by other authorities beyond the local district's control. Dealing with 20 districts in almost as many states makes uniform patterns of operation difficult; however, local district operation norms must be considered. An example was the conflict of project testing with existing testing programs in some states in terms of schedules of administration tests utilized, and conflicting mandates from state and federal agencies.

II. Recommendations Regarding Performance Contracting in General

1. It is recommended that districts pursuing performance contracting in any form determine their educational need as precisely as possible; determine to the best of their ability that they cannot fill the need with their own instructional resources; and *then* begin negotiations with a

contractor to fill the need. It is not recommended that districts enter contracts in order to avoid their responsibility in dealing with a difficult problem.

2. It is recommended that district personnel be utilized in the instructional process as much as possible.
3. Proper lead time in preparation and planning for a contract-venture is essential. Training of personnel, involvement of the community, total district staffs as well as building staffs, involvement of teacher's associations at all levels within the state and local areas, and clearance by state educational offices - all are crucial factors in the success of the program.
4. Deserving special emphasis in terms of time requirements is the contract itself. Each of the project directors faced moments of concern in terms of contract interpretation. Educators are not lawyers, but must be aware that a loosely written contract with inadequate attention to legal definition may be a source of embarrassment should court action ensue. It is the project directors' recommendation that attorneys be engaged to draw up, negotiate and interpret contracts.
5. A particular source of contractual confusion revolves about definition of role and responsibility of the contractor and subcontractor (school district and private company). In contracts in which the subcontractor provided his own instructional

personnel a source of ill feeling existed in authority relationships. It is recognized that problems of this nature are never likely to be totally resolved through means of contract, but verbal or unwritten agreements are totally inadequate.

6. It is recommended that criterion referenced tests be used as an evaluative base, insofar as they are available. Standardized tests, even though agreed upon by both contractor and subcontractor, can be questioned in terms of validity for this particular purpose and therefore constitute a potential for eventual dispute.
7. It is strongly recommended that any performance contract program be made an intergral part of the regular school program.
8. Perhaps the most universally agreed upon recommendation of the part of the project directors is that a district which considers performance contracting should be aware that personnel constitute the key to success. The strongest of curricular systems can be no better than the personnel operating them. It is agreed that if any of the subcontractors had felt their programs to be "people proof" they were less sure at the termination of the year.
9. Those who consider a performance contract should be reminded that the ultimate responsibility for the behavior and for the education of the child is that of the school and of the school board - not of the performance contractor.

10. It is recommended that representatives from contracting companies have proven administrative ability and experience, teaching experience at the appropriate level, and skill in interpersonal relations.

11. It is recommended that contracts specify that materials be on hand and that penalty clauses be a part of the contract.

12. It is recommended that on going evaluation be specified and criteria for acceptable performance be defined with contract cancellation if minimal performance lines are not maintained at established check points.

13. It is recommended that provision be included in all contracts for transition of programs into totally district operated ventures (turnkey).

In summary, the project directors feel that the Office of Economic Opportunity is to be commended for its willingness to enter a field of controversy in the hope of providing some answers to current questions regarding performance contracting. We feel that the problems encountered were in many cases inevitable, and better handled within the experimental context than in situations in which controls were not available. Through this year's experience, several approaches to the teaching of reading and mathematics skills were employed, with particular emphasis on diagnostic and prescriptive methods.

In virtually all centers, cases were seen in which students responded who for the most part had not responded in established programs. However, in all districts there were also problems of student control. The end result in terms of student achievement gain remains to be seen.

The essence of this experiment has been the relationship between public and private enterprise in the operation of our schools. We feel, as a result of the year's experiment, that both private enterprise and educators have gained in respect for one another and in understanding the complexities of public education.

The project directors, in conclusion, wish to comment on the concept of "accountability" as it relates to this project and to some of the problems faced in education today. No one can debate the desirability of accountability in the schools, but we, as project directors and educators, are concerned that this experiment not be caught in a web that it did not weave. This program was designed as an exercise in accountability, and as such requires an acceptable, definable function.

The skills of reading and mathematics are two educational areas which probably are most acceptable and definable. Even these, however, are by no means universally defined. Every educator struggles with the changing definitions and norms implied in such terms as "grade level."

The present means used to measure educational success to a level that a fair accountability requires are standarized tests, normed on the basis of average scores of large numbers of students. These scores are not absolutes, nor are they constant from norming period to norming period.

The tests, no matter how valid and reliable, are based on assumptions regarding the subject matter to be measured and are not necessarily based on the specific objectives of any single reading or mathematics program; the procedures and approaches vary considerably. As a result one reading or mathematics approach could very possibly be favored or ignored to a significant degree by a given standarized measure.

School districts throughout the nation are presently working toward the first step in a meaningful system of accountability - the precise definition of instructional objectives. This, in turn, demands a series of decisions on the part of school districts as to relative values of instructional matter and emphasis to be placed thereon. To date the lack of objectives and related value decisions has placed public education in a position of trying to do all things for all people - while increasing financial strictures make the task less and less possible.

When agreement has been reached on the objectives of our schools at all levels and when tests based on these objectives are precisely defined and available to measure success in these agreed upon skills - then accountability will be meaningful.

Progress is being made in both these areas - in defining objectives, and in creating criterion - referenced tests. This program in performance contracting can be an exercise in accountability only insofar as the state of the art has been perfected.

Bill Baker
Jacksonville

Doug Barnard
Mesa

Preston T. Bishop
Las Vegas

Dr. William P. Booth
Fresno

Ernest Cermola
Hartford

Hugo Ciullo
Philadelphia

Hollie Crawford
Stockton

Nelson Harris
Bronx

Dr. Ed. Ignas
Athens

Donald Olson
Anchorage

Lowell Pugh
Selmer

Fred Rotzler
Taft

Paul Steele
Portland

William Sternberg
Rockland

Sam Spaght
Wichita

William L. Tobias, Jr.
McComb

Dr. Don Waldrip
Dallas

Joan M. Webster
Grand Rapids

Bob Wulfestad
Seattle

227

Chapter VI

CONTRACTORS' STATEMENT

CONTRACTORS' STATEMENT

In order to respond to earlier interpretations and conclusions published by the Office of Economic Opportunity and its testing and analysis contractor, Battelle Memorial Institute, four of the six companies involved in the OEO performance contracting experiment[1] have concurred in this joint statement reflecting their views of the experiment and its results.

The contractors believe that, from its inception, elements of the experiment were so poorly conceived and conducted, particularly in its provisions for testing and evaluation, that these deficiencies should raise serious questions within the educational community on the broad generalized conclusions released by the OEO. The limited time for proposal submission, contract negotiations, school-contractor familiarization, program start-up, and over reaction to concerns about "teaching to the test" plagued the experiment throughout.

The situation which the companies were confronted with in the experiment can be illustrated by analogy to a hypothetical experiment to determine improvement in a particular athletic skill. Assume that the purpose of such an experiment was to compare a new method with the traditional approach of improving the athletic skill of high jumpers and that the simple objective of the experiment was to determine which approach would be more successful. Assume further, that the participants are divided into Group A to test the new method and Group B to test the traditional method. Finally, assume that, from time to time during the course of the experiment, the following events occur:

---

[1] Alpha Learning Systems, Inc , Learning Foundations, Inc , Plan Education Centers, Inc., and Singer/Graflex, Inc.

-- Participants who are selected for the program had been in training for three years.

-- Participants who had attained sufficient skill to high jump an average of 2.0 feet are placed in Group A and those with an average of 3 0 in Group B.

-- An arbitrary assumption is imposed, without consultation with or concurrence of the proponent of the new method, that all participants have sufficient skill to high jump at least 2.0 feet, even though 50% of the participants in Group A and 25% of the participants in Group B could only high jump 1.5 feet.

-- An arbitary rule is imposed, without consultation with or concurrence of the proponent of the new method, that the lowest level of the crossbar for the test of level of skill at the end will be 4.0 feet and that the improvement of skill of any participant who does not clear the crossbar at 4.0 feet was to assumed to be 0.

Under such a situation, it would be difficult to really determine what the level of improvement of each group was and almost impossible for Group A to achieve results better than Group B. Those participants whose beginning skill level was actually 1.5 feet would have to improve by more than 2.5 feet to reflect any gain at all. With a 3.0 feet actual gain, only .5 feet would be reflected. This gives an advantage to Group B because Group A had twice as many participants at this low level. Those participants whose beginning skill level was actually 4.0 or better would have the full gain reflected. This also gives Group B the advantage because of the overall higher beginning level of Group B. Group A participants could actually attain a 4.0 feet gain and Group B at 3.0 feet gain but yet have the conclusions reflect a 2.0 feet gain for Group A and a 1.875 feet gain for Group B.

The contractors' receipt of pre-test achievement scores confirmed the prevalent concern among the contractors that the testing was not going to provide for valid measurement of the effect of performance contracting on the reading and math skills of the disadvantaged students in the experimental groups. The array of test scores appeared to display an inconsistency with what was understood to be the levels determined for assignment by the OEO to groups at the onset of the program. Comparision of pre-test and assignment test scores was not possible as assignment test scores were not made available, although requested. At that point in time, halfway through the program, retesting to more adequately determine program entry levels was also not possible. Most important of all, in 17 of the 18 sites of the experiment, the average pre-test level of the control groups was significantly higher than that of the experimental group.

After receipt of pre-post test scores in late August, analyses by the contractors and their test consultants revealed the same inconsistencies ab-erved earlier and led the contractors to disagree with a number of conclusions by the OEO and Battelle. Among the many issues raised by the contractors are the following basic questions:

- -- Are experimental group vs. control group comparisons valid under the conditions imposed in the experiment?
- -- Are judgments about instructional programs accurate when tests used were not matched to instructional content?
- -- What effect did failure to administer appropriate test levels have in judging program effectiveness?
- -- Should criterion-referenced interim performance tests be categorically dismissed?
- -- Do "rate of learning" increases provide a more valid comparison of progress than comparison of actual scores?
- -- Does a one year experiment offer sufficient time to obtain summative conclusions?

In information disseminated to date regarding the research design, instrumentation, and analysis of outcomes, the OEO has consistently stated the conviction that it had sponsored a definitive experimental evaluation of the educational effectiveness of performance contracting among the disadvantaged -- and found performance contracting unequal to the task.

Moreover, official documents describing the project and summarizing its outcomes evidence a conscious effort to anticipate and forestall criticism related to the methodological aspects of project design, measurement, and data analysis. It should be noted that, unlike most government sponsored projects which seek to evaluate the effectiveness of an educational program, the sponsoring agency assumed, directly or through its agents, full responsibility for the research design, for instrumentation, for data collection, for data analysis, and for interpretation of results. To all intents and purposes the OEO functioned in the performance contract project not as a sponsor but as a research institute which delegated only the instructional responsibilities to the performance contractors.

The reason for adverting to these facts is to place in perspective the relation between the OEO and the performance contractors with respect to the issues to be raised below. The OEO has alleged, for example, that the performance contractors' "agreement to be judged on the basis of standardized tests was an indication of their belief in the validity of the tests".2/ It would be more accurate to say that it was an indication of the belief of the contractors that the OEO could and would identify and choose standardized tests that would constitute a fair basis for payment and tests which would provide a valid basis for evaluation of instructional outcomes. In other words,

---

2/ Summary of Preliminary Results, OEO pamphlet 3400-5, pg. 14.

the performance contractors agreed, for purposes of payment, to live with whatever validity the test selected might possess. They did not thereby contract to forego a reasonable retrospective concern for the appropriateness of the tests, nor did they contract to accept a partnership share of the responsibility for the suitability of the tests chosen by the OEO. The OEO's complete and total responsibility for the actual suitability of the measures for evaluating the attainment of project objectives is a natural consequence of its appropriation of absolute authority over every aspect of the evaluation process.

Quite aside from the matter of payment itself--which is obviously a matter of no small concern to the contractors--is the issue of the scientific integrity of the conclusion that performance contracting does not work. The conclusion is based upon the finding that children receiving remedial instruction under contractor auspices failed to exhibit substantial gains or to exceed control group performances on standardized tests of general educational achievement. It is the conclusion, not the findings, which is being questioned here. The only direct and appropriate measure of the effectiveness of instruction is the learning criterion, not its correlates, i.e., general measures of achievement or "school success", as the OEO preliminary report suggests.

The most fundamental question that can be raised with repsect to any research project is the relevance of the kind and quality of the evidence collected, in this case test scores, to the purposes of the investigation.

The relevance of the general educational achievement of disadvantaged children in the areas of reading and math is, in of itself, a matter of unquestionable importance. Whether general educational achievement as defined by standardized test performance is, should be, or in this project was the instructional goal of services purchased from performance contractors is an important question. "A performance contractor signs an agreement to improve students' performance in certain basic skills by set amounts" (emphasis added).[3] General measures of education achievement do not measure basic learning skills or basic knowledges, either in toto or, which is more to the point, in their separate subject area subtests; they measure instead a wide range of highly complex skills in somewhat cursory fashion. This is exemplified by the fact that it is not at all infrequent that an improvement of a raw score by no more than ten items will result in a full year gain in grade-equivalent scores. To achieve sensitivity to even substantial changes in basic functional deficiencies which plague the disadvantaged learner one simply cannot justify the scatter gun approach of the general achievement test.

Moreover, by basic skills, one ordinarily means those reading and computing skills prerequisite to progress towards the complex objectives typical of classroom instruction. It is neither logical nor realistic, therefore, to expect immediate transfer of learning to result from instruction in basic skills.

---

[3] OEO pamphlet 3400-5, pg. 2.

The function of remediation is restoration of the capability to profit from classroom instruction. It is reasonable to expect improved response to classroom instruction and improvement of performance on general achievement tests subsequent to remedial skills training; it is irrational to expect such improvement as a naturally occurring concomitant of remediation.

The inference is inescapable that from the standpoint of content validity, standardized measures of general educational achievement, unless related to the content and format of a particular instructional program, do not constitute acceptable measures of the extent to which performance contracting, or any instructional program, succeeds or fails in the attainment of remedial basic skills training with the disadvantaged.

By the same token, it is difficult to appreciate the rationale for the criticism of criterion-referenced interim performance tests on grounds other than content validity. The fact, for example, that "less than 1% of the children failed to answer at least 75% of the questions correctly" and that, therefore, they were "too easy"[4] is curious psychometric logic--unless, of course, one is interested more in the measurement of individual differences among children than in measuring what each child knew or learned. Even a test on which every child answers every question correctly would not ipso facto be too easy, provided the test could claim content or curricular validity and provided the children had not been somehow coached in the specifics of test content. Such a result might simply mean that the children had, in fact, been taught effectively. The presumption that a test accompanied by such results is too easy appears to represent an indirect and implicit apriori rejection of the effectiveness of performance contracting, i.e., if the criterion-referenced tests appear to support the effectiveness of instruction, they must have been too easy.

---

[4] OEO pamphlet 3400-5, pg. 16.

A number of more specific questions concerning the efficacy of standardized achievement tests, which have not been matched to program content, for the evaluation of performance contracting outcomes are of considerable substantive importance. The use of grade equivalent scores at all, let alone as the principal basis for evaluation, was unfortunate at best. Despite their popularity and despite the seeming interpretative simplicity of grade equivalents, the use of such tests is fraught with statistical and interpretative pitfalls. The use of grade equivalence to assess the extent of a pupil's performance relative to actual grade placement is deceptively uninformative. Differences between grade equivalent scores and the actual level of grade placement are not only unreliable but indefensible as representations of the developmental progress they seem to suggest. The methods by which different grade equivalent scores are obtained within each grade level, i.e., 6.0 - 6.9, bear no relation to the actual developmental progress in scholastic achievement.

Furthermore, measured differences which cut across grade levels, i.e., 6.5 - 7.5, vary in their meaning from one grade level to the next. In other words, there is no reason to believe that grade equivalent gains at one level, i.e., +1.0 from a grade equivalent of 6.5 to one of 7.5, represents the same amount of progress as a gain at some other level, i.e., 5.0 - 6.0, 5.5 - 6.5, etc.

The educational level of the children to whom a test is administered is a matter of obvious and essential importance in the selection of an appropriate test for the assessment of both achievement standards and achievement gains. Aside from questions of content validity already raised, another and equally serious concern is the selection of an appropriate level of difficulty for the children being tested. This means that tests must be selected which measure achievement within the actual range of the functional skills possessed by the persons tested. When one tests educationally disadvantaged children, this creates a readily understandable problem. Their functional level of skill in areas measured by standardized achievement tests is known to be appreciably below the level of children from those segments of the population on whom such tests are standardized originally. Instruments were used which were designed to the grade in which the students were even though the contract specifically stated that only students with grade level deficiencies would be eligible for the experiment. This resulted in purported test results far beyond any arguable range of reliability of the level of test used.

According to testing done by the performance contractors, it appears that a significant number of students were at a low enough level to reflect a fictitious pre-test experiment level based on statistical probability alone. Such students could, of course, have a real gain of 2 to 3 grade levels and yet show no gain at all because of the fictitious beginning level resulting from use of instruments which cannot reliably test at levels as low as that of these students. The selection of test levels almost assured invalidity of any conclusions reached on the project from the outset. For example, an 8th grader who scored 7.0 or higher on the pre-test was not qualified for the project and any 8th grader who scored much below 7.0 was not on a level within the range of reliability for that particular test instrument.

In very simple terms, and aside from technical considerations of reliability, the probable result of miscalculating the test level appropriate to the testees' functional achievement level was to examine them on skills which, qualitatively and quantitatively speaking, they did not possess. By the same token it becomes difficult, if not impossible, to specify with precision what skills they do possess. Achievement measurement is the assessment of what one knows, not what one does not know. This problem is most strikingly exemplified in performance observed on one of the tests administered to the first graders. The test contained an extensive set of minimal performance screening items which, to all intents and purposes, made no contribution to grade equivalent scores.

A child could, and children did, perform so poorly on pretests that very substantial gains on screening-item performance from pre-test to post-test would not result in score improvement by so much as one-tenth of a "grade level".

The contractors feel that a comparison based on improvement in learning rates would be the most appropriate, especially in view of the fact that the control group, except for one, were composed of students who appeared to have a higher learning rate than the experimental groups. Obviously, if one student in the 6th grade is at a 1.5 level (or a .25 learning rate) and another is a 3.0 level (or a .50 learning rate) a significant difference in achievement is attained if both progress 1.0 in the 3rd grade. The rate of learning for the first student is 400% of his historical attainment whereas that of the second student is 200% of his historical attainment. Preliminary analysis of such data as has been made available to the contractors indicates that some experimental groups may have done significantly better than the control groups on the basis of a comparison of gains in learning rates.

Since conclusions based exclusively on test results regarding the effectiveness of performance contracting have been widely disseminated and publicized, it is informative to note the temporal relationship of those conclusions to empirical evidence concerning the technical adequacy of the tests. These announcements of findings and conclusions preceded the actual investigation of the tests' reliability for project participants, in spite of the evaluation contractors' frank expression of concern for test reliability in a population so distinctively different from the population for which the test was constructed and on which it was standardized and normed.[5]

Finally, although this commentary is most directly concerned with measurement issues that affect the interpretation of project outcomes, it is also important to refer at least in general to the matters involving sampling design and the statistical analysis appropriate thereto. Difficulties involved in the design of field experiments notwithstanding, the absence of randomization of pupil assignment to experimental and control groups may not be lightly dismissed. The failure to effect such randomization constitutes a substantive and significant departure from the essential definition of a true experiment. Whatever the magnitude of a study, random allocation to experimental and control groups from a common pool of available subjects remains the only scientifically dependable method of neutralizing the influence of irrelevant extraneous factors upon criterion performance. The evaluation contractor's avoidance of statistical analyses which assume randomization is commendable. But it must also be pointed out that the use of complex methods of regression analysis[6] does not ameliorate the inherent weakness of a design which necessitates their use; it merely acknowledges and accommodates that weakness.

---

[5] Battelle Interim Report, pg. 61.
[6] Battelle Interim Report, pg. 62-73.

There is simply no known way, statistical or otherwise, to prevent the confounding influence of extraneous factors from producing differential effects upon the criterion performance of non-randomized groups. There is, of course, no certitude that confounding actually will occur under such condition. The problem is that there is no assurance that it will not. And if it should occur, there is no precise method for identifying its specific source or magnitude. From one point of view it might be said that the analytic methods employed were those best suited to the kind and quality of the data collected. It is not too harsh to say, however, that what this means is that the analytic methods used were the least objectionable under the circumstances of the sampling design.

In summary, the performance contracting project cannot realistically be described as a definitive, rigouous experimental investigation of the impact of performance contracting in the remediation of basic learning skills or educational achievement among the disadvantaged in general. It was actually a very large quasi-experiment, of limited external validity, fraught with start-up difficulties, teacher resistance, poor testing conditions, and other problems that adversely affected the experimental groups. Apart from the testing and evaluation inconsistencies, limitation of the experiment to a one-year life term was a serious mistake. It is conservatively estimated that the first four months were devoted to reaching the normal September status for experimental students. Concurring with the need for a second year for testing of the educational innovations introduced by the contractors, many of the school districts exerted efforts to find funding to maintain the programs a second year.

The point must be made that had the contractors known that the control groups would not be randomly matched with the experimental groups, had they known that improper levels of achievement tests would be used, and that the tests would not be matched with the instructional programs, the contractors would never have entered into the OEO performance contracting experiment under such terms.

The disheartening thing that the contractors feel is unwarranted about the conclusions drawn from the experiment is the increased polarization between the educational community and the private sector just at the time when educational technology has reached a stage of development that can produce significant benefits for American education. Private companies have produced rather startling gains working with disadvantaged youth and adults in tutoring centers and manpower programs, and believe that the private sector makes a contribution to public education in America if it can work in full cooperation with, and not in opposition to, the existing school systems.

Issues such as those described in this statement have made the contractors involved in performance contracting conclude that at best the results are inconclusive. However, the experiment was not without value. A number of concerns of those interested in the impact of new technology in the classroom have been identified and perhaps clarified. Emphasis has been given to measurement and the use and misuse of achievement tests. Many sweeping generalizations can be put to rest; quick cures, and short-range demonstrations alike, can be deemed inappropriate to the magnitude of the task.

Finally, it is the recommendation of the contractors that the base established by this experiment be built upon for further investigation. Accountability, by performance contracting or other means, should proceed under controlled experimentation and measurement.

☆U.S. GOVERNMENT PRINTING OFFICE: 1972 722-314/476 1-3

Distribution: S